OPEN SCIENCE MUNI SEMINAR

WORKSHOP: ANONYMIZATION OF RESEARCH DATA

2021-03-12

MUNI

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

PROJECT: CZ.02.2.69/0.0/0.0/18_054/0014703

M U N I

Workshop: Anonymization of Research Data

Fabian Praßer (Berlin Institute of Health, Universitätsmedizin Berlin)
Michal Růžička, Jiří Marek (Institute of Computer Science, Masaryk University)

2021-03-12

Workshop Schedule

1. Keynote Lecture: Fabian Praßer
2. Questions from the Audience
3. Next Steps: Michal Růžička, Jiří Marek

MUNI

MUNI

1. Keynote Lecture

MUNI

2. Questions from the Audience

Questions from the Audience (1/4)

1. Even with anonymized quantitative data, the combination of data can lead to potential identification – how to handle it?
2. Do all participants have to agree to the publication of data? If so, how is such consent demonstrated when showing signed informed consent would result in a breach of anonymity?
3. Despite anonymization, participants can recognize themselves – what to do about that?
4. Data retention – how to store completed printed questionnaires?
5. How to store and share sensitive, quantitative data?
6. I currently work with clinical data and some of them we plan to anonymize.
7. I am interested in anonymization of qualitative data, for example (semistructured) interviews.

MUNI

Questions from the Audience (2/4)

8. I would be particularly interested in anonymizing interviews (unstructured, narrative, relatively long, where participants tell their life story).
9. We are interested in anonymizing photographs taken by research participants and on which persons could appear (we ask participants who take pictures in the school environment so that persons do not appear there, but they can still be there).
10. The problem of anonymizing clinical population data has been puzzling statisticians for a long time. I'm honestly interested in what today's science has to offer, whether any available software tools can help with data anonymization.

MUNI

Questions from the Audience (3/4)

11. How to evaluate the anonymization of full-text documents?
    Machine anonymization of full-text documents only recently became plausible. Most research papers focus on the first part of the problem: recognizing named entities related to personal data (e.g., IberLEF MEDDOCAN 2019 [1]), which is straightforward to evaluate. However, I have found only a few papers on the second part of the problem: evaluating anonymization's quality and strength for the recognized sensitive data. For example, the models t-pat (Anandan 2011 [2]) and C-sanitized (Sanchez 2016 [3]) base their evaluation on measuring semantic similarity with Information Content, which is an outdated concept. In (Hassan, Sanchez 2019 [4]), where they use Word2Vec embeddings instead, they did not address evaluation at all, so it seems that they did not solve the problem. This leads me to believe that an evaluation model usable with modern NLP methods like word/subword embeddings does not exist yet. I am very interested in your thoughts on the topic.
    [1] http://ceur-ws.org/Vol-2421/
    [2] https://ieeexplore.ieee.org/abstract/document/6040853
    [3] https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.23363
    [4] https://ieeexplore.ieee.org/abstract/document/8887419

MUNI

Questions from the Audience (4/4)

12. Do you know of any real-world deployments of autonomous machine anonymization of full-text documents?
    So far I've only read about assistive deployment, where the sensitive terms are highlighted but the final decision and anonymization is left to a human expert.
13. I plan research using pair statistical tests in longer-term – repeated trials with a similar group of people over several years. How to pseudo-/anonymize the data while preserving me interconnect the same persons' answers in separate research trials?

MUNI

MUNI

3. Next Steps

Legal Aspects and Support at MUNI

1. Open Science legal consultations
2. Preparation of follow-up workshops on the legal issues of anonymization
   - GDPR article 89 and its use in science
   - Longitudinal studies

MUNI

Technical Aspects and Support at MUNI

1. Development of SensitiveCloud infrastructure.
   - OpenStack based environment for secure processing of sensitive data.
2. Preparation of follow-up workshops on the use of anonymization tools.
   - Singapore PDPC Guide (https://muni.cz/go/5a1665)
   - ARX – Data Anonymization Tool (https://arx.deidentifier.org/)
   - NLM-Scrubber (https://scrubber.nlm.nih.gov/)
   - Automated Defacing Tools mri_deface (https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface)

MUNI

OPEN
SCIENCE
MUNI
SEMINAR

THANK YOU
FOR YOUR
ATTENTION!

2021-03-12

MUNI

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

PROJECT: CZ.02.2.69/0.0/0.0/18_054/0014703