National Coordination Point
Research Data Management

Pitch:

Achieving Anonymization in Human-Derived Research Data

Developing clear definitions and assessing re-identification risks for specific data types

Name: E. Jessica Hrudey
Date: 11 March, 2019
Contact details: e.j.hrudey@vu.nl or research.data.fgb@vu.nl

Goal:
The GDPR does not apply to personal data that has been anonymized. If human-derived research data can be anonymized, they can be more readily shared for the purposes of open science.
Unfortunately, the current definition for anonymized data according to the GDPR is vague and open to interpretation; anonymized data is information that "does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable". Assurance of anonymity requires an assessment of "all the means reasonably likely to be used…to identify the natural person directly or indirectly" as well as "whether means are reasonably likely to be used to identify the natural person, (and) account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of processing and technological developments".

This ambiguous definition creates uncertainty for researchers about whether any human-derived data can ever be considered anonymous and it places a heavy burden upon researchers to consider all possible ways in which the data may not be anonymous. This is especially complex in a world where so much data about individuals can now be found online and where machine learning algorithms may be able to identify individuals based solely on their writing style.

The aim of this pitch is to develop a clear definition for anonymous human-derived research data and to inventory the re-identification risks of various data types. These results will guide future pitches on this topic by providing a clear and concrete framework for developing:

1. education programs about anonymization and;
2. methods to assess the effectiveness of anonymization software and tools

Cross-Institution Importance:
This issue is important across the Netherlands and the EU because open science can and should be achieved, but this must be done within the requirements of the GDPR. Anonymized data can be shared readily, but if data are not considered sufficiently anonymous under the GDPR, such

data sharing could be considered a data breach. Because the GDPR's definition of anonymous data is so vague, it creates uncertainty about whether *any* human-derived data can be considered anonymous. Institutions may respond by handling all human-derived data as personal data under the GDPR, no matter how low the risk of re-identification, however this will lead to an excessive administrative burden upon institutions and researchers who must also manage these data on the long-term for open science purposes.

If data can be reasonably considered anonymous, the GDPR no longer applies, thereby reducing this administrative burden. Clearly many types of human-derived research data can only be pseudonymized and must still be carefully managed as required by the GDPR, however, if a portion of research data can be considered anonymous, the total burden upon institutions and researchers will be lower.

Currently there are differing opinions between institutions about how strictly the Dutch Data Protection Authority will enforce the definition of anonymous data. Therefore, bringing together stakeholders from a variety of institutions will allow for a discussion and hopefully an agreement on what is a reasonable definition of anonymous human-derived research data. As well, to inventory the privacy risks of various data types, researchers, data scientists, anonymization experts and privacy lawyers must come together to discuss the likelihood, feasibility and impact of re-identification with these various types of human-derived research data.

Deliverable & outline:

- A definition of anonymous research data
    - In future phases of the proposed pitch, this definition may be used to develop legally binding advice for researchers
- An inventory of several types of human-derived research data and the associated risks for re-identification
    - The task group will initially focus on data types that are generally thought to be anonymous (e.g. skull-stripped MRIs, reaction time data, accelerometry, other highly variable physical measurements) so that researchers who use these types of data have a resource for assessing re-identification risk in data that may be otherwise assumed to be anonymous
    - If time permits, the task group will inventory the re-identification risks of other human-derived data types that are generally considered indirect identifiers

Expertise needed:

- Data scientists
- Privacy lawyers
- Researchers
- Anonymization experts

Duration:
Beginning of April through End of July.

Previous work and what is missing:
A variety of anonymization tools exist and there are many publications on statistical theory regarding K-anonymity, L-diversity and other concepts.
There is also a fantastic online course from a research group in Canada about anonymizing data and the senior investigator of that research group co-authored a white paper recommending that the anonymization of data under the purview of the GDPR is ideal for both the data users and the people from whom the data are collected.
There remains, however, a lack of uncertainty, particularly from privacy law, as to whether anonymization techniques and tools will meet the requirements of the GDPR. There is room for interpretation within the text of the GDPR and because there are no legal cases regarding insufficient anonymization since the enforcement of the GDPR, it is unclear how strictly the Dutch Data Protection Authority will interpret the GDPR text.
Additionally, there is a lack of specific information about the risks of re-identification within various types of data. The GDPR requires an extensive assessment of the risks of re-identification before data can be considered anonymous; a summarized overview of several data types and the re-identification risks will ease the burden on researchers who would otherwise have to determine this information by themselves. Ultimately, researchers need to know how they can share data effectively, safely, but also without excessive administrative burden.

Without clearer definitions and insight into the risks of re-identification, sharing data will remain a complex issue.