Landelijk Coördinatiepunt
Research Data Management

PITCH

Anonimisering van mensgebonden onderzoeksdata

Ontwikkelen van heldere definities en het vaststellen van re-identificatierisico's voor specifieke typen data

Naam: E. Jessica Hrudey
Datum: 11 maart 2019
Contactgegevens: e.j.hrudey@vu.nl of research.data.fgb@vu.nl

Doel:

De AVG geldt niet voor geanonimiseerde persoonlijke data. Als data uit mensgebonden onderzoek geanonimiseerd kan worden, is het delen hiervan in het kader van Open Science een stuk makkelijker.

Helaas is de huidige definitie van geanonimiseerde data in de AVG wat vaag en open voor interpretatie: geanonimiseerde data is informatie die "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". Garanties voor anonimiteit vragen een assessment van "all the means reasonably likely to be used…to identify the natural person directly or indirectly" en ook "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". Deze ambigue definitie zorgt ervoor dat onderzoekers onzeker zijn of, en wát voor mensgebonden/persoonlijke data ooit beschouwd kan worden als anoniem. Het legt een zware druk op onderzoekers om alle opties in kaart te brengen om inzichtelijk te krijgen wanneer data mogelijk niet anoniem is. Dit is vooral ingewikkeld in een wereld waarin zo ontzettend veel data over individuen online gevonden kan worden en waarin machine learning algoritmes zelfs in staat lijken te zijn om individuen te identificeren op basis van alleen maar hun schrijfstijl.

Het doel van deze pitch is om een heldere definitie te ontwikkelen voor anonieme mensgebonden onderzoeksdata en om een inventarisatie te maken van de re-identificatierisico's van verschillende typen data. Deze resultaten zullen in een vervolgpitch moeten leiden tot een helder en concreet raamwerk voor de ontwikkeling van:

1. Trainingprogramma's over anonimisering;
2. Methoden om de effecitiviteit van anonimiseringssoftware en – tools te kunnen meten

Instituutsoverstijgend belang
Dit onderwerp is van belang in Nederland en in de EU omdat open science alleen maar bereikt kan worden binnen de vereisten en kaders van de AVG. Geanonimiseerde data lijkt goed deelbaar, maar zodra die data onder de AVG wordt beschouwd als “niet voldoende anoniem”, leidt het delen ervan tot een datalek.
Omdat de definitie van anonieme data in de AVG zo vaag is, speelt de vraag of mensgebonden data überhaupt wel als anoniem kan worden beschouwd. Instituten kunnen hierop reageren door alle mensgebonden data als persoonlijke data onder de AVG te beschouwen, ongeacht het werkelijke risico op re-identificatie. Dit zal onvermijdbaar leiden tot een excessieve administratieve last voor instituten en onderzoekers die deze data moeten managen en beheren voor de lange termijn open science-doelen.

Als data redelijkerwijs beschouwd kan worden als anoniem, dan is de AVG niet langer meer van toepassing en kan de administratieve last worden verminderd. Het is duidelijk dat veel typen mensgebonden onderzoeksdata alleen maar gepseudonimiseerd kunnen worden en dat het nog steeds noodzakelijk is deze voorzichtig te behandelen, zoals de AVG vraagt. Maar: als een deel van de onderzoeksdata beschouwd kan worden als anoniem, dan wordt de last voor onderzoekers en onderzoeksinstellingen een stuk minder.

Momenteel leven er verschillende opvattingen tussen de instituten over hoe strikt de Autoriteit Persoonsgegevens de definitie van anonieme data zal opvatten. Door stakeholders van verschillende instituten bij elkaar te brengen, kan er ruimte ontstaan voor discussie en hopelijk overeenstemming over een werkbare definitie van anonieme mensgebonden onderzoeksdata. Door de privacyrisico’s van verschillende typen data te kunnen inventariseren is het van belang om onderzoekers, data scientists, anonimiseringsexperts en privacy-juristen bij elkaar te brengen om waarschijnlijkheid, haalbaarheid en impact van re-identificatie met verschillende typen mensgebonden onderzoeksdata vast te kunnen stellen.

Resultaat:

- Een heldere definitie van geanonimiseerde onderzoeksdata
    - In een vervolgpitch zou deze definitie kunnen dienen om juridisch bindend advies voor onderzoekers te ontwikkelen
- Een inventaris van verschillende typen mensgebonden onderzoeksdata en de bijbehorende risico’s voor re-identificatie
    - De taakgroep zal beginnen met typen data die in het algemeen beschouwd worden als anoniem (bijvoorbeeld skull-stripped MRIs, reaction time data, accelerometry, other highly variable physical measurements) zodat onderzoekers die werken met deze typen data een bron hebben re-identificatierisico’s te toetsen
    - Als er genoeg tijd is, zal de taakgroep een inventaris maken van het re-identificatierisico van andere mensgebonden typen van data die algemeen beschouwd worden als indirecte identifiers.

Gevraagde expertise:

- Data scientists
- Privacy-juristen
- Onderzoekers
- Anonimisering experts

Looptijd Taakgroep
Begin april – eind juli

Wat is er al bekend en wat mist nog?
Er is een grote variatie aan anonimiserings-tools beschikbaar en er zijn verschillende publicaties beschikbaar over statistische theorieën over K-anonimiteit, L-diversiteit en andere concepten.

Er bestaat ook een fantastische online cursus van een onderzoeksgroep in Canada over anonimsering van data. De senior-onderzoeker van die onderzoeksgroep was medeauteur van een white paper waarin wordt aanbevolen dat anonimisering van data, bezien vanuit het perspectief van de AVG, ideaal is voor zowel de dataonderzoekers, als voor de datasubjecten.

Het blijft echter onzeker, vooral op basis van de privacy wetgeving, of technieken en tools voor anonimisering wel voldoen aan de eisen van de AVG. De tekst van de AVG biedt ruimte voor interpretatie en omdat er sinds de AVG geen juridische cases zijn over onvoldoende anonimisering is het onduidelijk hoe strikt de Nederlandse Autoriteit Persoonsgegevens de AVG zal interpreteren.

Daarnaast is er ook gebrek aan specifieke informatie over de risico's van re-identificatie van verschillende typen data. De AVG vraagt om een uitgebreid assessment van de riscio's van re-identificatie voordat data als anoniem beschouwd kunnen worden.
Een samenvattend overzicht van de verschillende typen data en de bijbehorende re-identificatierisico's zullen de druk op de onderzoekers verminderen. Ze hoeven deze informatie niet meer zelf te achterhalen. Uiteindelijk moeten de onderzoekers weten hoe ze effectief en veilig data kunnen delen, zonder dat ze daar een extensieve administratie voor bij hoeven te houden.

Zonder werkbare, heldere definities en inzicht in de risico's van re-identificatie, zal data delen een complex onderwerp blijven.