ANONIMISERING van DATA SETS

Voorstel voor het verantwoorden van het anonimiseren van data[1]

Nanda Piersma, december 2021

Introductie

Het omgaan met gevoelige data vereist de uiterste zorgvuldigheid van de onderzoeker. Het betreft het verzamelen, het opslaan, het toegang geven, het bewerken en het archiveren van de gegevens. Onder de privacywetgeving (AVG, GDPR) is het verboden om persoonsgegevens op te slaan, te gebruiken en door te geven zonder daarvoor een wettelijke grondslag te hebben. Voor onderzoek wordt in beginsel uitgegaan van de grondslag toestemming. Ook herleidbare persoonsgegevens vallen hier onder, deze bestaan uit schijnbaar anonieme gegevens die door een combinatie van kenmerken tot een specifieke persoon leiden. Een voorbeeld is een combinatie van locatiegegevens met tijdsaanduidingen binnen een school, die eenvoudig (met een schoolrooster) zijn te herleiden tot een specifieke docent. Het anonimiseren van gegevens is een bewerking die de persoonsgegevens onleesbaar maakt of verwijdert. Een succesvolle anonimiseringsbewerking maakt het onmogelijk om personen te identificeren in de gegevens in een dataset. Let wel, goed anonimiseren is niet eenvoudig en de aanpak hiervoor moet per situatie worden bepaald aan de hand van bestaande richtlijnen https://www.lcrdm.nl/nieuws/handreiking-anonimisering-beschikbaar .

Bij het verzamelen van data ontstaat een zogenaamde “ruwe” dataset. Als de set betrekking heeft op mensen, bevat de set vaak persoonsgegevens. Dit zijn gegevens waarmee een (natuurlijk) persoon geïdentificeerd of identificeerbaar is. Het is raadzaam om uit de ruwe dataset een onderzoeksset en/of werkset te maken. Deze set bevat alleen de relevante gegevens voor de analyse. In bijna alle gevallen is dit een anonieme set.

Hieronder geven wij een aantal richtlijnen over hoe om te gaan met anonimisering van persoonsgegevens in datasets. Deze richtlijnen zijn van toepassing voor het onderzoek aan de Hogeschool van Amsterdam. Ze zijn een onderdeel van de ethische richtlijnen en datamanagementprocedures. Ook zijn deze onderdeel van het proces waarmee de digitale veiligheid van de gegevensverwerking wordt gewaarborgd. Zie https://az.hva.nl/medewerkers/staven-en-diensten/az-lemmas/medewerkers/hva-breed/juridische-zaken/privacy/modellen/modellen-2.html (onder punt 5).

Er zijn meerdere aspecten rondom het anonimiseren die zorgvuldig moeten gebeuren, en ook goed beschreven moeten zijn.

1. Anonimiseren of pseudonimiseren van een dataset
2. Gebruik van geanonimiseerde datasets
3. Verantwoorden van het anonimiseren

[1] Dit voorstel is vanuit de ethische commissie opgesteld en bedoeld als handreiking voor onderzoekers die een ethische toetsing aanvragen en de datastewards van de faculteiten van de Hogeschool van Amsterdam

1. Anonimiseren van een dataset

Persoonsgegevens kunnen worden geanonimiseerd óf gepseudonimiseerd:

a. Het verwijderen van de namen of persoonsaanduiding. Elk datapunt is dan onherleidbaar tot een individu. Als een persoon meerdere keren in de dataset voorkomt, dan is dat niet meer zichtbaar. Let hierbij op dat de combinatie van datapunten toch herleidbaar is tot een persoon. Bepaal daarom per geval de methode waarop gegevens worden geanonimiseerd. Zie https://www.lcrdm.nl/files/lcrdm/2020-01/LCRDM%20Risicomanagement%20voor%20data%20over%20mensen.pdf en mogelijk https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_nl.pd. Deze bewerking is niet mee terug te draaien en heet ***anonimiseren***.
b. Het coderen van de namen of persoonsduidingen. Elk individu krijgt een eigen code, dus elke keer dezelfde code voor eenzelfde persoon. De manier waarop de code wordt toegewezen kan verschillen, de meest gebruikte techniek maakt gebruik van een hash tabel, waarmee een sleutel, of code, wordt geassocieerd met een datawaarde (zoals een naam). De meeste kwaliteitssoftware of programmeertalen kennen een hashalgoritme. Omdat de hash tabel techniek omkeerbaar is, wordt deze techniek niet anonimiseren genoemd, maar ***pseudonimiseren.*** Het gebruiken van pseudonimiseren lijkt soms voldoende om de privacy te beschermen, maar deze is met een sleutel ongedaan te maken en dat moet verantwoord worden. Als dezelfde onderzoeksresultaten ook met anonimisering kunnen worden bereikt, ben je verplicht de werkset te anonimiseren.

2. Gebruik en opslag van geanonimiseerde datasets

Er zijn vier fases met bewerkingen op een datasets.

Verzamelfase: het bouwen van de ruwe dataset MET persoonsgegevens.
Bewerkingsfase: het bouwen van de geanonimiseerde of de gepseudonimiseerde dataset met sleutel.
Analysefase: activiteiten om vanuit de bewerkte dataset de openbaar beschikbare -+onderzoeksresultaten te halen.
Archieffase: bewaarperiode voor verantwoording van het afgeronde onderzoek.

Een onderzoek kan alleen plaatsvinden met toestemming van de deelnemers, vastgelegd in een toestemmingsprocedure. Daarin worden deelnemers vooraf geïnformeerd over het onderzoek en geven zij toestemming voor de medewerking en voor het gebruiken van hun persoonlijke data. Omdat mensen de mogelijkheid moet worden geboden om de toestemming in te trekken, zal de ruwe dataset tijdens het onderzoek beschikbaar moeten blijven. Wanneer een persoon is verwijderd uit de ruwe dataset dan zal het onderzoek herhaald worden zonder deze persoon, inclusief de anonimisering. Dit moet ook duidelijk zijn in de toestemmingsprocedure en de informatiebrieven.

Het omgaan met de datasets is per fase anders.

2.1 Verzamelfase

De ruwe dataset is eigendom van de onderzoekers. Bij gebruik van externe partijen (software leveranciers of technische partijen) zijn er afspraken over eigenaarschap en toegang. Tot de ruwe data (met persoonsgegevens) hebben andere partijen alleen toegang als dat echt noodzakelijk is. Bij voorbeeld om te koppelen met CBS-microdata. De ruwe dataset wordt op een beveiligde plaats bewaard en is alleen toegankelijk voor de onderzoekers.

2.2 Bewerkingsfase

Er wordt een nieuwe, bewerkte dataset gebouwd met alleen de kenmerken nodig voor het onderzoek en waarin de persoonsgegevens zijn gemaskeerd (anonimisering of pseudonimisering). Ook voor deze dataset worden afspraken gemaakt over toegang en de software waarmee de bewerkingen plaatsvinden. Het eigenaarschap van de bewerkte dataset is opnieuw van de onderzoekers. De bewerkte datasets kunnen op verzoek met andere onderzoekers worden gedeeld voor verificatie of vervolgonderzoek, mits geanonimiseerd, of gepseudonimiseerd ZONDER sleutel.

2.3 Analysefase

Tijdens de analysefase wordt de vertrouwelijke informatie omgezet in conclusies en onderzoeksresultaten die met een groot publiek worden gedeeld. Als er in deze fase ook datasets worden gemaakt en gedeeld, dan zullen deze moeten voldoen aan alle eisen van de AvG en geen enkele privacy (herleidbare) gegevens moeten bevatten. Deze datasets kunnen in openbare omgeving worden gedeeld, en zonder toestemming worden gebruikt door derden.

2.4 Archieffase

Na afloop van het onderzoek worden alleen de bewerkte dataset, en de openbaar beschikbaar gestelde datasets uit de analysefase bewaard. Het onderzoek kan daarmee worden herhaald en geverifieerd. De persoonsgegevens zijn daarvoor niet nodig. De ruwe datasets (met de persoonsgegevens) worden daarom vrijwel altijd vernietigd, enkele uitzonderingen nagelaten (bijvoorbeeld medische gegevens die in een andere context bewaard moeten blijven).

3. Verantwoorden van de anonimiseringsbewerkingen

In het datamanagement plan moet duidelijk worden

1. welke persoonsgegevens in de (ruwe) dataset staan,
2. welke bewerkingen worden uitgevoerd op de data om deze te pseudo/anonimiseren,
3. bij pseudonimisering: welke sleutel er is om te de-anonimiseren (bijvoorbeeld een logboek, of een hashtabel), en uitleg waarom anonimisering niet kan.
4. Wie de data anonimiseert, of hoe, omdat deze persoon of software toegang moet hebben tot de ruwe data met de privacygevoelige gegevens,
5. Welke gegevens wel worden opgenomen in de bewerkte, anonieme dataset,
6. Waarom er geen sprake is van herleidbare persoonsgegevens met de wel vermelde gegevens in de geanonimiseerde dataset,
7. De (opslag)plaats van elke dataset gedurende en na afloop van het onderzoek.

Conclusie

Het uiteindelijke doel van het onderzoek is om algemene inzichten te verkrijgen uit het bestuderen van data met persoon gerelateerde gegevens. Deze individuen willen hun medewerking beloond zien met een goed beschermde omgeving en zorgvuldige informatie over het gebruik van hun gegevens. Voor het werken met datasets met persoonsgegevens zijn hier een aantal richtlijnen gegeven voor de stappen en de verantwoording van de anonimisering.

Verder lezen

Privacy:
algemene informatie AVG
https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/algemene-informatie-avg/algemene-informatie-avg

HvA informatie Privacy
https://az.hva.nl/medewerkers/dmci/az-lemmas/medewerkers/hva-breed/juridische-zaken/privacy/privacy.html?origin=mVz26ZjITCqUsiP4wOc5Hg

Data management protocollen
Research data management:
https://az.hva.nl/medewerkers/dmci/az-lemmas/medewerkers/hva-breed/bibliotheek/onderzoeksdata/onderzoeksdata.html?origin=mVz26ZjITCqUsiP4wOc5Hg

Ethische commissie
https://az.hva.nl/medewerkers/dmci/az-lemmas/medewerkers/hva-breed/oo/onderzoek-ethiek-en-integriteit/onderzoek-ethiek-en-integriteit.html?origin=mVz26ZjITCqUsiP4wOc5Hg