Dienst Uitvoering Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

Directie
Onderwijsinstellingen

Afdeling
Informatieproducten / Team Synthetische Data

Contact
synthetische.data@duo.nl

memo

Privacy en Juridische Aspecten Synthetische tabulaire data

Managementsamenvatting
DUO maakt sinds 2022 synthetische tabulaire data producten. De afdelingen Informatieproducten en Compliance hebben samengewerkt om de juridische en privacy aspecten hiervan te verhelderen en vast te leggen. In de kern: 1) Synthetische tabulaire data maken het mogelijk om data te delen die originele data in hoge mate benadert, zonder daarbij persoonsgegevens vrij te geven. 2) De privacyrisico's die synthetische tabulaire data met zich meebrengen, zijn vergelijkbaar met die van geaggregeerde data.

Inleiding

In dit memo wordt uiteen gezet wat de wens is van DUO rondom privacybeleid op synthetische tabulaire data. Deze uiteenzetting is tot stand gekomen na afstemming over dit onderwerp tussen Compliance en ontwikkelaars van synthetische tabulaire data binnen DUO. In dit memo wordt de definitie van synthetische tabulaire data besproken en wordt beschreven hoe er vanuit een juridisch en privacyrechtelijk perspectief naar synthetische tabulaire data gekeken kan worden.

Wat zijn synthetische tabulaire data?

Synthetische tabulaire data zijn tabulaire data die kunstmatig gegenereerd worden.[1] Tabulaire data zijn gegevens gerepresenteerd in tabelvorm, vaak gevestigd in de vorm van een bestand of enkele tabel in een database. Het doel van synthetische tabulaire data is om het mogelijk te maken om data te leveren die originele data in hoge mate benadert, zonder daarbij persoonsgegevens vrij te geven. De mogelijkheden van toepassing van synthetische tabulaire data zijn breed, maar vooralsnog ligt de nadruk op het faciliteren van onderzoek en het testen van software. Synthetische tabulaire data kunnen oplopend in representativiteit van originele data gegeneerd worden door middel van globale statistieken betreffende de originele data en machine-learningtechnieken, op basis van onderliggende statistische samenhang van variabelen, of een combinatie van beide. Het is belangrijk om synthetische tabulaire data te onderscheiden van andere technieken. Synthetische tabulaire data (zoals beschreven in deze memo) betreffen in geen geval gepseudonimiseerde data, ongeacht de complexiteit van versleuteling, gemaskeerde data, data op basis van beslisregels, data op basis van een simulatie of data gebaseerd op basis van random-sampling. Het gebruik van synthetische tabulaire data voor het faciliteren van onderzoek binnen DUO is verder beschreven in de bijgevoegde Factsheet Synthetische data.

[1] Alvaro Figuera & Bruno Vas, Survey on Synthetic Data Generation, Evaluation Methods and GANs, 2022; César Augusto Fontanillo López en Abdullah Elbi, 'On the Legal Nature of Synthetic Data', 2022.

Synthetische tabulaire data en de AVG

Voor het maken van synthetische tabulaire data zijn datasets met persoonsgegevens nodig. Dit betekent dat het maken van synthetische datasets een verwerking van persoonsgegevens betreft en dus binnen de reikwijdte van de AVG valt. Dit heeft tot gevolg dat je voor het maken van een synthetische dataset een legitiem doel moet hebben en je moet kunnen baseren op één van de grondslagen als bedoeld in artikel 6 AVG. De verwerking van persoonsgegevens voor het maken van een synthetische dataset valt onder artikel 6 lid 1 onder de AVG. Het gaat hier om een verwerking in het kader van het uitoefenen van een taak van algemeen belang. DUO heeft als taak: *het verstrekken en beheren van onderwijsinformatie ten behoeve van beleid en onderwijsveld.*[2] Het gaat hier om een zogenaamde verdere verwerking (dus anders dan het oorspronkelijke verzameldoel). "*De verdere verwerking met het oog op archivering in het algemeen belang, wetenschappelijk of historisch onderzoek of statistische doeleinden, moet als een met de aanvankelijke doeleinden verenigbare rechtmatige verwerking worden beschouwd."*[3]

De resulterende, volledig gesynthetiseerde dataset, daarentegen bevat geen persoonsgegevens meer en valt buiten de reikwijdte van de AVG. *"Synthetic data is not real data about a person. (…), a single record in a synthetic dataset does not correspond to an individual or record in the original (real) dataset. (…) It does not include an identifier that corresponds to an actual natural person. It does not reference the physical, physiological, genetic, mental, economic, cultural, or social identity of an actual natural person. In short, a fully synthetic dataset does not meet the GDPR definition of "personal data.""*[4]

Ook het onderzoekscentrum van de Europese Commissie stelt dat synthetische tabulaire data geen persoonsgegevens bevat: *"As opposed to real data, synthetic data are de-personalized (not personal) data, and therefore can be used in cases in which the target is not to identify a certain person."*[5]

Wanneer data niet volledig of niet op een juiste manier gesynthetiseerd worden, kan het voorkomen dat synthetische tabulaire data niet volledig anoniem zijn, waardoor je eindigt met een pseudonieme synthetische dataset.[6] Bij een pseudonieme dataset is er nog wel sprake van persoonsgegevens en moet deze dataset dus overeenkomstig de AVG verwerkt worden. Om dit te voorkomen hebben we binnen DUO kaders ontwikkeld waarin de definitie en voorwaarden van gebruik bij synthetische data zo zijn vastgelegd dat door DUO gesynthetiseerde datasets altijd alleen anonieme gegevens bevatten.[7] Hierbij is dus in geen geval een 1-op-1 relatie te ontsluiten.

---

[2] Organisatie- en mandaatbesluit OCW 2008, bijlage 1, hoofdstuk 9, https://wetten.overheid.nl/BWBR0023543/2022-10-15#Bijlage.

[3] Overweging 50 AVG

[4] Khaled el Emam, Lucy Mosquera en Richard Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data, First edition, Boston, 2020, hfst 6.

[5] European Commission Joint Research Centre, Multipurpose Synthetic Population for Policy Applications, LU: Publications Office, 2022, p. 17.

[6] "(…) according to the identifiability test as enshrined in the definition of personal data, synthetic data will be considered pseudonymous or anonymous data depending on the appropriateness of the data synthesis and the related ex-post control mechanisms." (López en Elbi 2022).

[7] Kader Statistische Beveiliging en Pilotverslag Synthetische Data

Tot slot is het goed te vermelden dat het feit dat synthetische tabulaire data geen persoonsgegevens meer bevatten tot gevolg heeft dat het delen, openbaar maken of aanbieden van synthetische datasets niet door de AVG beperkt wordt en data van DUO voor een breder publiek en voor een grotere verscheidenheid aan doeleinden kan worden gebruikt. Daarnaast wijst de Europese Commissie op andere mogelijkheden, zoals het verrijken van bepaalde karakteristieken van synthetische datasets.[8]

De vergelijking tussen synthetische tabulaire data en geaggregeerde data
De restrisico's van synthetische tabulaire data zijn vergelijkbaar met die van geaggregeerde data.

Geaggregeerde data worden in de AVG in beginsel als anonieme gegevens aangemerkt, op voorwaarde dat de data niet tot op de persoon herleidbaar zijn.[9] Geaggregeerde gegevens zijn belangrijk voor het bepalen van bijvoorbeeld overheidsbeleid en de AVG onderschrijft de publieke belangen die met statistische analyse gediend kunnen worden. Denk bijvoorbeeld aan statistische analyse door het Centraal Bureau voor de Statistiek. Dit is essentieel voor op informatie gebaseerde beleidsvorming door de overheid. Als gebruik wordt gemaakt van statistische geaggregeerde data gelden de regels voor het verwerken van statistische gegevens, die normen inhouden voor onder meer vertrouwelijkheid en veiligheid.

Lopez & Elbi nemen als uitgangspunt dat geaggregeerde data, waarbij informatie niet te herleiden is naar een individu, buiten het Europese databeveiligingsraamwerk moeten vallen.[10] Voor synthetische tabulaire data geldt hetzelfde: bestaande personen hebben tot gegevens in de synthetische dataset geleid, maar zijn niet als zodanig te identificeren. Deze zouden daarom hetzelfde behandeld moeten worden als geaggregeerde data.

Een verschil tussen synthetische tabulaire data en geaggregeerde data is dat er bij synthetische data sprake is van een probabilistische relatie tussen individuen en de dataset. Dat betekent concreet dat het zowel voor buitenstaanders als voor de data-eigenaar niet te identificeren is welke bestaande personen er tot een gegeven in de synthetische tabulaire data heeft geleid. Voor geaggregeerde data geldt dit alleen voor buitenstaanders.

Binnen het voorgestelde informatieclassificatiebeleid van OCW dat opgesteld wordt kan gesteld worden dat synthetische tabulaire data in de regel – net als geaggregeerde data - binnen niveau 0 van de informatie- en dataclassificatie valt. Dit niveau omvat data zonder persoonsgegevens. Dit kan geconcludeerd worden uit het feit dat de data niet herleidbaar is naar bestaande personen.

Bij het toetsen van privacy binnen synthetische tabulaire data wordt het kader Statistische Beveiliging in acht genomen. Privacy wordt gewaarborgd doordat synthetische tabulaire gegevens worden gegenereerd met behulp van de wetenschappelijk onderbouwde stand der techniek en doordat

8 *"(…) it enables the possibility to enhance certain characteristics that might not be cut by the real data, e.g. outliers, biases, etc."* EC Joint research centre, 2022, p.17.

9 Bart van der Sloot, Sascha van Schendel, en César Augusto Fontanillo López, 'The influence of (technical) developments on the concept of personal data in relation to the GDPR', Tilburg Institute for Law, Technology and Society (TILT), 2022, p.40.

10 César Augusto Fontanillo López en Abdullah Elbi, 'On the Legal Nature of Synthetic Data', 2022.

controlemechanismen op het gegenereerde product worden uitgevoerd. Hierbij wordt rekening gehouden met de observatie data naarmate de analysewaarde hoger is (dus hoe preciezer de originele data nagebootst wordt), het risico dat privacy in het geding kan komen ook hoger is.

Het vergelijkbaar behandelen van synthetische tabulaire data ten opzichte van geaggregeerde data heeft een aantal praktische gevolgen. Zo moeten dezelfde normen van vertrouwelijkheid worden toegepast, met name betreffende de groepsinformatie, de onomkeerbaarheid van het syntheseproces en de voorwaarden aan publicatie van algoritmen. Daarnaast geldt dat andere overwegingen die gemaakt worden over het publiceren van een geaggregeerde dataset – zoals de mogelijke impact – op dezelfde manier gemaakt moeten worden voor de analoge synthetische dataset.

Conclusie
Bij het maken van synthetische tabulaire data worden persoonsgegevens verwerkt als bedoeld in de AVG. Je moet voor het maken van een synthetische dataset dus een doel en (wettelijke) grondslag hebben. Maar, synthetische tabulaire data (zoals wij dat definiëren) vallen zelf niet onder de AVG als het eenmaal (correct) gemaakt is. Daarbij is het essentieel dat er geen 1-op-1 relatie is met bestaande personen en dat de synthese onomkeerbaar is. Ondanks dat synthetische tabulaire data niet onder de AVG valt zijn er nog steeds 'normen van vertrouwelijkheid en veiligheid' van toepassing zoals bij geaggregeerde data. Informatieproducten vraagt daarom accordering van deze stellingname en communicatie hierover.