UWV ICT Richtlijn Data Anonimisering en Pseudonimisering

Versie 1.1

Opdrachtgever

Naam: Alex Kooistra

Functie: CISO

Datum: 27-07-2018

Versiebeheer

Versie	Datum	Doelgroep	Info
0.1	Juli 2016	Werkgroep TDA richtlijn, TSC en coördinatoren IB&P.	Notitie R. Foorthuis als basis nemen voor op te stellen richtlijn, beslisboom toevoegen. Notitie uitgezet bij coördinatoren IB&P en TSC.
0.2	Sept 2016	Werkgroep TDA richtlijn	Tekstuele/redigerende aanpassingen.
0.3	Okt 2016	Werkgroep TDA richtlijn, TSC en coördinatoren IB&P.	Notitie uitgezet bij coördinatoren IB&P en TSC.
0.4	Nov 2016	coördinatoren IB&P.	Aanpassingen n.a.v. versie 0.3
0.9	Dec 2016	AB	Akkoord met kleine aanpassingen
1.0	13-12-2016	IV-team en Tactisch IB&P overleg.	Akkoord
1.01	27-6-2018	Intern CISO-office	Herziene conceptversie i.v.m. invoering AVG
1.02	06-07-2018	coördinatoren IB&P en TSC.	richtlijn ter review bij TSC.
1.03	12-07-2018	Tactische coalitie	Notitie ingebracht bij tactische coalitie ter instemming
1.04	18-07-2018	IV team	Vernieuwde bijlage richtlijn BZ classificatie ingevoegd; richtlijn ter instemming aan IV team
1.1	24-07-2018	Intranet beleidspublicatie	Definitieve versie na goedkeuring IV-team

Inhoud

1 Aanleiding

De General Data Protection Regulation (GDPR), de Algemene Verordening Gegevensbescherming (AVG), is vanaf 25 mei 2018 van toepassing in de hele EU. Vanaf die datum vervangt de AVG de Wet bescherming persoonsgegevens (WBP), die sinds 1 september 2001 van kracht is.

De AVG stelt als één van de eisen dat privacy by design en privacy by default (PbDD) bij nieuwbouw of vernieuwing van bestaande van producten en diensten, processen en systemen moet worden toegepast. De bedoeling is om al in de ontwerpfase privacybescherming van de betrokkene mee te nemen ('by design'), en tevens standaard de meest privacybeschermende instellingen toe te passen ('by default').

In het "Beleid Privacy by Design en by Default, Privacybescherming in het ontwerp en standaard instellingen (PBDD-beleid)" geeft UWV invulling aan de vereisten vanuit de Algemene Verordening Gegevensbescherming (AVG) om in een zo vroeg mogelijk stadium van de gegevensverwerking de afwegingen te maken ten aanzien van de inbreuk op de persoonlijke levenssfeer (privacy) van de betrokkene waarvan de persoonsgegevens worden verwerkt. Het PBDD-beleid geeft richting voor uitwerking in toepasbare en uitvoerbare principes, richtlijnen en maatregelen waarmee ten aanzien van privacy by design en by default aan de AVG kan worden voldaan.

In dit document worden twee maatregelen in het PBDD-beleid, nader uitgewerkt:
- Anonimiseren
- Pseudominiseren

De wettelijke tekst wordt door de Autoriteit Persoonsgegevens (AP) als volgt uitgelegd:

- "Privacy by design houdt in dat u als organisatie al **tijdens de ontwikkeling** van producten en diensten (zoals informatiesystemen) ten eerste aandacht besteedt aan **privacyverhogende maatregelen**, ook wel privacy enhancing technologies (PET) genoemd[1]. Ten tweede houdt u rekening met **dataminimalisatie**: u verwerkt zo min mogelijk persoonsgegevens, dat wil zeggen alleen de gegevens die noodzakelijk zijn voor het doel van de verwerking.
- Privacy by default houdt in dat u **technische en organisatorische maatregelen** moet nemen om ervoor te zorgen dat u, als **standaard**, alléén persoonsgegevens verwerkt die noodzakelijk zijn voor het specifieke doel dat u wilt bereiken.

Het nieuwe PBDD-beleid heeft een beperkte invloed op de bestaande richtlijn Test Data Anonimisering uit 2016. Vanuit het perspectief van de AVG is het van belang om expliciet onderscheid te maken tussen geanonimiseerde en gepseudonimiseerde gegevens. Bij volledig geanonimiseerde gegevens is privacywetgeving, zoals de AVG, niet meer van toepassing. Bij de gepseudonimiseerde data is dit wel het geval.

1.1 Wijzigingen n.a.v. herziening IB&P richtlijn OTAP procedure

Dit herziene beleid heeft geleid tot genuanceerde wijzigingen van de UWV ICT richtlijn Data Anonimisering en Pseudonimisering.

De belangrijkste wijzing is dat er nadrukkelijk onderscheidt wordt gemaakt tussen de begrippen anonimisering en pseudonimisering. De drie soorten anonimisering zoals in de bestaande richtlijn werd gehanteerd is gewijzigd naar gebruik van pseudonimisering en anonimisering.

[1] https://autoriteitpersoonsgegevens.nl/sites/default/files/downloads/technologie/witboek_pet.pdf

2 Inleiding

UWV verwerkt gegevens en moet daarbij aan verschillende wet- en regelgeving, en samenwerkingsafspraken voldoen. De belangrijkste regels voor de omgang met persoonsgegevens in Nederland zijn vastgelegd in de Algemene Verordening Gegevensbescherming (AVG).De richtlijnen en maatregelen en maatregelen in dit document zijn een concrete uitwerking van de volgende AVG-verplichting:

Artikel 5 AVG lid 1f: Persoonsgegevens moeten door het nemen van passende technische of organisatorische maatregelen op een dusdanige manier worden verwerkt dat een passende beveiliging ervan gewaarborgd is, en dat zij onder meer beschermd zijn tegen ongeoorloofde of onrechtmatige verwerking en tegen onopzettelijk verlies, vernietiging of beschadiging ("integriteit en vertrouwelijkheid

De Richtlijnen zijn ook toepasbaar bij overige wet- en regelgeving en samenwerkingsafspraken waar anominisering of pseudominisering wordt vereist.

Deze versie van het beleid is alleen aangepast op grond van het inwerking treden van de AVG per 25 mei 2018. De AVG onderscheid mogelijkheden van maatregelen zoals anonimisering, pseudonimisering, dataminimalisatie, encryptie, beperking toegangscontrole, logging en monitoring en verwijdering/vernietiging van gegevens teneinde zo min mogelijk persoonsgegevens te verwerken, dat wil zeggen alleen de gegevens die noodzakelijk zijn voor het doel van de verwerking.

2.1 Doel

Deze richtlijn heeft tot doel een passend niveau van privacy- en informatiebeveiliging te garanderen bij de verwerking van persoonsgegevens voor secundair gebruik, zoals testprocessen, data-analyses en leveringen aan derde partijen, zodat wordt voldaan aan wet- en regelgeving (o.a. AVG, BIR, SUWI) en aan samenwerkingsafspraken waaraan het UWV is verbonden.

2.3 Uitgangspunt

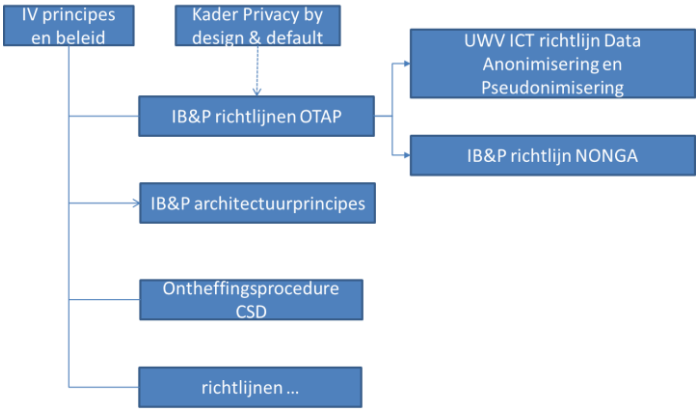
UWV moet voldoen aan informatiebeveiligings- en privacybeschermingsnormen (zoals AVG, BIR en SUWI) die aan wet- en regelgevingen aan samenwerkingsafspraken verbonden zijn. Zij eist van de applicatieverantwoordelijken, dat voor secundair gebruik van gegevens de volgende regels worden gehanteerd:

- Persoonsgegevens worden geanonimiseerd, waar dit mogelijk en haalbaar is.
- Er worden geen persoonsgegevens verwerkt, waarvoor geen rechtmatig verwerkingsdoel is bepaald.
- Dataminimalisatie zal moeten worden toegepast. Gegevens die voor een beoogde verwerking als zodanig overbodig zijn, zijn dus niet zijn toegestaan, en zullen moeten worden gewist.
  Voorbeelden daarvan zijn het overbodig gebruik van het BSN-nummer of gebruik van gegevens van VIP's.
- Door dataminimalisatie worden alleen die gegevens beschikbaar gesteld, die voor het verwerkingsdoel noodzakelijk zijn.
  -Dit betekent dat data die niet voor het doel van de verwerking van belang is wordt verwijderd, en dat de inbreuk op privacy op de overgebleven persoonsgegevens, met behulp van anonimiserings- en pseudonimiseringstechnieken zo veel mogelijk wordt geminimaliseerd.

Dit geldt voor testomgevingen (t.b.v testen van applicaties), analyseomgevingen (t.b.v. data-analyses), als ook voor de acceptatie en productieomgeving waarbij persoonsgegevens voor secundaire doeleinden worden verwerkt en voor het doorgeven van persoonsgegevens naar andere partijen.

2.4 Toelichting IB&P-richtlijnen

Hieronder is de ICT richtlijn Data anonimisering en pseudonimisering en de plaats in de beleidsboom schematisch weergegeven.



Figuur 1 – Plaats IB&P- en ICT richtlijnen in beleidsboom UWV

Deze richtlijn geeft invulling aan de vereisten uit de volgende wet- en regelgeving, en samenwerkingsafspraken:

1. Telecommunicatiewet
   -specifiek is Artikel 11.5 van belang: Anonimisering verkeers- en rekeninggegevens
2. Algemene verordening Gegevensbescherming (AVG)
3. Uitvoeringswet AVG (UAVG)
   -Specifiek is van Artikel 46 van belang: Verwerking nationaal identificatienummer.
   Een nummer dat ter identificatie van een persoon bij wet is voorgeschreven, wordt bij de verwerking van persoonsgegevens slechts gebruikt ter uitvoering van de desbetreffende wet dan wel voor doeleinden bij de wet bepaald.

Overigens nog de opmerking dat het gebruik van BSN gericht is op de doelbinding voor verwerking. Binnen UWV wordt daarmee zowel de directe als de indirecte doelbinding bedoeld.

2..5 Beveiligingstechnieken

De AVG vereist dat bij de verwerking van persoonsgegevens passende technische of organisatorische worden genomen. De term Privacy Enhancing Technologies (PET) wordt gehanteerd om alle ICT-middelen aan te duiden die gebruikt kunnen worden om persoonsgegevens te beschermen.

'Privacy enhancing technologies' (pet) is de verzamelnaam voor een aantal technieken die de verantwoordelijke kan toepassen om bij het verwerken van persoonsgegevens de risico's voor de betrokkenen te beperken.
Een centraal principe van 'pet' is het verminderen of opheffen van de herleidbaarheid: dit is de mate waarin persoonsgegevens kunnen worden herleid tot de betrokkenen. De zwaarste vorm van pet is anonimisering van de verwerkte persoonsgegevens.

UWV heeft in haar begroting ruimte gereserveerd ten behoeve van dergelijke technische IB&P maatregelen. Ook bij deze maatregelen dient door de verantwoordelijken een risico afweging te worden gemaakt en valt hiermee onder het comply or explain principe.

Bij het Test Service Center (TSC) beschikt men over de ICT-middelen DATPROF waarin aantal Privacy Enhancing Technieken in een omgeving zijn samengebracht en geïntegreerd, als hulpmiddel bij anonimiserings- en pseudominisatieprocessen. Via de TCS-dienstverlening is deze tool als generiek hulpmiddel beschikbaar, inclusief ondersteunende diensten.

Hiermee verwordt de dienst Test Data Anonimisering met de tool DATPROF die door het TSC wordt uitgevoerd voor UWV, tot een generieke technische IB&P maatregel bij het testen met productie-data. Naast reeds bestaande oplossingen in deze, bij andere partijen.

DATPROF wordt alleen ingezet voor de maken van statische datasets. Dit is een batch-georiënteerde aanpak voor anonimisering en pseudonimisering. Het biedt geen oplossing voor datasets die dynamisch, of op basis van een real-time gegevensopvraag, geanonimiseerd of gepseudonimiseerd moeten worden.

Van anonimisering is sprake als de gegevens op geen enkele manier meer tot de betrokkene te herleiden zijn. Er is dan geen sprake meer van persoonsgegevens en de AVG is niet meer van toepassing op de gegevens.

Een lichtere vorm van pet is het scheiden van de verwerkte persoonsgegevens in (zeer goed beveiligde) identificerende gegevens en niet-identificerende gegevens. Alleen met behulp van de identificerende gegevens kan de identiteit van de betrokkenen worden achterhaald.

2.6 Omschrijving en Definities

In het PBDD-beleid is opgenomen, ten aanzien van anonimiseren:

Wanneer is vastgesteld dat persoonsgegevens worden verwerkt, moet worden bepaald of anonimisering van persoonsgegevens kan worden toegepast. Met de "UWV ICT Richtlijn Test Data Anonimisering" (Nu "UWV ICT richtlijn Data Anonimisering en Pseudonimisering") kan dit worden bepaald. Geanonimiseerde persoonsgegevens worden door de AVG niet meer als persoonsidentificeerbaar beschouwd en in dat geval zijn verder verdere maatregelen niet nodig. De afwegingen en de keuze om de persoonsgegevens volledig te anonimiseren moet wel worden vastgelegd. Dit geldt zowel voor de ontwikkel, test, acceptatie-omgevingen, als ook in de productieomgeving en bij leveringen.

Ten aanzien van pseudonimiseren is in het PBDD-beleid opgenomen:
Wanneer anonimiseren en dataminimalisatie niet toegepast kunnen worden, of er blijft nog een hoog restrisico over, kan Pseudominiseren worden toegepast. Bij Pseudominiseren worden persoonsgegevens op een zodanige wijze verwerkt dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat aanvullende gegevens (de toegepaste pseudonimiseringssleutel) worden gebruikt. Deze aanvullende gegevens (sleutels) moeten dan wel apart worden bewaard, en technische en organisatorische maatregelen moeten worden genomen.

Omdat gepseudonimiseerde gegevens persoonsgegevens zijn, is de AVG van toepassing.

Encryptie (versleuteling) en pseudonimisering komen op hetzelfde neer. Met de juiste sleutel of aanvullende gegevens kunnen de (persoons)gegevens ontsleuteld en leesbaar gemaakt worden, waardoor herleiden tot individuen weer mogelijk is. Het verschil is dat versleutelde data (of encrypted data) geheel onleesbaar is zonder de juiste sleutel. Gepseudonimiseerde data is leesbare data, waarbij identificerende gegevens zijn verwijderd of vervangen, waarbij de herleidbaarheid van de betrokkene weliswaar wordt verminderd maar niet volledig teniet wordt gedaan. Er is hierbij sprake van (de mogelijkheid tot) omkeerbaarheid.

Anonimisering daarentegen is onomkeerbaar, waarbij na toepassing ervan herleiden van gegevens tot individuen niet meer mogelijk is. Vormen hiervan zijn randomisatie en generalisatie.

Anonimiseren:
Met anonimiseren worden de gegevens omgezet naar "een vorm die identificatie van de betrokkene feitelijk niet langer mogelijk maakt.
Van anonimisering is sprake als de gegevens op geen enkele manier meer tot de betrokkene te herleiden zijn. Er is dan geen sprake meer van persoonsgegevens en de AVG is niet meer van toepassing op de gegevens (CBP, richtsnoer 2013).

Dit houdt in dat niemand meer tot herleiding in staat is.

Pseudonimiseren
In dit document spreken we van pseudonimiseren als de gebruiker van de geanonimiseerde dataset de gegevens niet langer naar de betrokkenen in kwestie kan herleiden, maar de beheerder van deze set dit nog wel kan (bijvoorbeeld met geheime sleutels).

Het bieden van een passend beveiligingsniveau d.m.v. het voorkomen van herleidbaarheid voor deze datagebruiker is namelijk het doel van dit document. Bedenk echter dat zolang de beheerder de geheime sleutels heeft, de gegevens voor de organisatie nog wel herleidbaar zijn. Juridisch gezien is er dan nog steeds sprake van persoonsgegevens en zijn de gegevens op organisatieniveau dus niet geanonimiseerd.

Pseudonimiseren is het vervangen van een identificerend gegeven door een ander identificerend gegeven, waarbij de herleidbaarheid van de betrokkene weliswaar wordt verminderd maar niet volledig teniet wordt gedaan (CBP, 2016). Een hacker kan dan mogelijk een manier vinden om de data via geavanceerdere technieken weer herleidbaar te maken. Dit speelt met name bij het vervangen van identificerende gegevens middels cryptografische technieken.

Pseudonimisering mag niet worden gezien als synoniem van anonimisering. Pseudonimisering beperkt alleen de koppelbaarheid van een dataset aan de oorspronkelijke identiteit van een betrokkene, en is bijgevolg een nuttige maatregel om gegevens te beveiligen.
Dit betekent dat pseudonimisering in samenhang met andere maatregelen voor informatiebeveiliging en privacybescherming moeten worden beschouwd en dat deze andere maatregelen niet kan vervangen.

3. Criteria bij anonimiseren / pseudonimiseren

Deze UWV ICT Richtlijn data Anonimisering en pseudonimisering presenteert criteria voor het pseudonimiseren of anonimiseren van gestructureerde datasets waarin persoonsgegevens zijn opgenomen ten behoeve van het gebruik binnen UWV.

De focus in dit document ligt op het pseudonimiseren / anonimiseren van nominatieve datasets, dus van microdata in plaats van statistisch geaggregeerde rapportages. Verder gaat de aandacht uit naar de functionele aspecten van anonimiseren, niet naar de tools die ervoor gebruikt kunnen worden.

Het is bij anonimiseren/pseudonimiseren van belang vooraf goed te analyseren welke typen datagebruik van belang zijn en wat de aard van de gegevens is.

De verschillende *typen datagebruik* (testen, data analyse, beheer, etc) brengen verschillende eisen met zich mee t.a.v. hoe en in welke mate de data geanonimiseerd dienen te worden. Voor het performancetesten van software is het bijvoorbeeld geen probleem als aan BSN's random een variatie van adressen wordt toegekend, terwijl dit voor data analyse ongewenst is omdat de representatieve structuur van de data daarmee wordt verminkt.

Ook binnen één type datagebruik kan sprake zijn van verschillende eisen. Voor een data analyse die zich richt op verklarend modelleren (het maken van causale statistische modellen) of anomaly detection (het identificeren van vreemde datapunten) is het bijvoorbeeld geen probleem als de namen van steden worden geanonimiseerd. Denk hierbij aan het vervangen van de oorspronkelijke namen van steden door fictieve steden, zoals Amsterdam door Gotham City. Echter, dit is geen optie voor een opdracht waarbij de data analist moet rapporteren over in welke steden arbeidsbemiddeling het meest effectief wordt uitgevoerd.

De *aard van de gegevens* speelt daarnaast ook een rol bij de te maken keuzes in het anonimiseerproces. Het is bijvoorbeeld van belang of de desbetreffende set statisch of dynamisch is, dus of deze niet, respectievelijk wel, wordt aangevuld met nieuwe data die kunnen verwijzen naar de reeds aanwezige gegevens. Daarnaast speelt mee in welke mate het mogelijk is om betrokkenen te identificeren door de gegevens set te koppelen met andere data. In de praktijk is dit vaak lastig in te schatten.

Op basis van de criteria in het adviesrapport anonimiseringstechnieken, van de Artikel 29-werkgroep (WP29), is anonimiseren in de praktijk niet altijd haalbaar. In de meeste gevallen zal het pseudonimisering en anonimisering proces in gepseudonimiseerde persoonsgegevens resulteren, waarvoor een passend beveiligingsniveau moet worden gegarandeerd.

4. Het anonimiseren van persoonsgegevens

In bijlage 1: 'Het identificeren van personen' wordt concrete input geboden voor de wijze waarop persoonsgegevens geanonimiseerd moeten worden. Anonimiseren dient alleen te gebeuren in situaties waarin er geen eenvoudiger mogelijkheid bestaat de privacy van data te waarborgen.

Een voorbeeld van zo'n eenvoudiger mogelijkheid is het simpelweg *niet* gebruiken van persoonsgegevens in testomgevingen, en hier gefingeerde testgevallen in te zetten.

Dit ligt voor functioneel testen ook voor de hand. Tevens is dit de juridisch gewenste situatie. Er zijn echter situaties waarin toch productiegegevens nodig zijn. Als het systeem getest moet worden op robuustheid tegen vreemde gevallen en uitzonderingssituaties, zullen deze in een grote productionele dataset gevonden moeten worden. Dit garandeert namelijk dat het om realistische situaties gaat. Bovendien is de realiteit altijd 'gekker' dan wij als mensen kunnen bedenken.

Met Performance- en stresstesten zijn op basis van de OTAP richtlijnen productiedata te gebruiken. Daarnaast is het altijd nodig data analyses op productiedata uit te voeren om bijvoorbeeld anomalieën (afwijkingen, opmerkelijkheden) op te sporen of trends te definiëren. Hierbij gaat het niet om de individuele gevallen. Kortom, er zijn diverse situaties waarin het anonimiseren van data relevant is.

Per situatie moet bekeken worden tot hoever de dataset 'verhaspeld' dient te worden zodat een voor die setting voldoende veilig niveau van anonimisering is bereikt (ICO, 2012).

Zo zal een dataset die buiten de organisatiegrens ter beschikking wordt gesteld (voor testen, onderzoek of andere doeleinden) geanonimiseerd moeten worden en zijn er mogelijkheden voor pseudonimisering bij gebruik van een een dataset die louter wordt gebruikt door enkele interne testers.

In het algemeen geldt het uitgangspunt dat goed anonimiseren lastig is (Sweeney, 2000). Terughoudendheid is gewenst met het zomaar ter beschikking stellen van geanonimiseerde datasets.

In bijlage 2 wordt de wijze van anonimiseren/pseudonimiseren verder uiteengezet.

5. Niveaus van anonimisering / pseudonimisering en aandachtspunten

Het staat vast dat elk van de bekende anonimiserings- en pseudominiseringstechnieken tekortkomingen vertonen. Dit neemt niet weg dat elke techniek in de gegeven omstandigheden en context geschikt kan zijn om het beoogde doel te verwezenlijken zonder de privacy van de betrokkenen in gevaar te brengen. De optimale oplossing dient daarom per geval te worden vastgesteld, zo nodig door verschillende technieken te combineren. Hierbij moet een afweging worden gemaakt tussen twee conflicterende doelstellingen:

- Is de gegevensverzameling bruikbaar en representatief is voor het beoogde verwerkingsdoel
- Is het risico van inbreuk op de privacy van betrokkenen tot een acceptabel niveau teruggebracht.

Ter ondersteuning van deze afweging worden in de volgende paragraaf een aantal pseudominisatie niveaus onderkend, aangevuld met aandachtpunten.

Men kan verschillende niveaus van anonimisering / pseudonimisering onderkennen:

- *Geen:*
  De gegevens worden niet geanonimiseerd of gepseudonimiseerd omdat ze worden gebruikt door medewerkers in de primaire bedrijfsvoering (zoals claimbeslissers en werkbemiddelaars) of andere gebruikers met doelbinding. (In deze richtlijn niet het object van onderwerp)of er is geen herleidbaarheid naar personen.

- *Laag:*
  De gegevens worden op de primaire en mogelijk een aantal secundaire identificerende sleutels geanonimiseerd. De gegevens worden hierbij niet of nauwelijks verminkt. Inhoudelijk vinden er geen transformaties of verwijderingen plaats, en er worden format-preserving hashes gebruikt. De gepseudonimiseerde set is toegankelijk voor een vooraf bepaalde kleine groep van gebruikers, zoals uitvoerders van performancetesten of data-analisten.. De gegevens staan in een afgeschermde omgeving waarbij passende informatiebeveiligingsmaatregelen van toepassing zijn. Specifieke maatregelen moeten ervoor zorgen dat de kans op re-identificatie, door analyses of koppelingen met andere gegevens, tot een acceptabel niveau wordt teruggebracht.

- *Middel:*
  De gegevens worden op de primaire en de secundaire identificerende sleutels geanonimiseerd. Daar waar bepaalde overige gegevens duidelijk informatie over individuele betrokkenen prijs geven, worden ze veralgemeniseerd of verwijderd. De gegevens worden in bepaalde mate verminkt. De gepseudonimiseerde set voor een vooraf bepaalde middelgrote groep UWV-interne gebruikers en/of een zeer beperkt aantal externe vertrouwde gebruikers.

- *Hoog:*
  De gegevens worden op de primaire en secundaire identificerende sleutels geanonimiseerd, evenals op alle kenmerkende velden die mogelijk identificerende informatie prijs zouden kunnen geven. Omdat het er geen risico's worden genomen, worden veel gegevens verwijderd of veralgemeniseerd. Daarnaast worden benodigde Risicoklasse III gegevens voor de zekerheid getransformeerd, zodat niet langer duidelijk is wat voor gegevens het betreft.

  Om de kans op re-identificatie tot een minimaal niveau terug te brengen (en daarmee een mogelijkheid van een datalek te voorkomen), kunnen daarnaast bijvoorbeeld fictieve locaties gebruikt worden ten behoeve van stedennamen. De data zijn fors verminkt en niet meer geschikt voor alle typen gebruikt. Een geanonimiseerde set is geschikt voor grootschalig extern gebruik, waarin UWV geen controle meer heeft op het gegevensgebruik. Omdat onherleidbaar anonimiseren in de praktijk moeilijk te realiseren is, is het advies om hier terughoudend en zorgvuldig mee om te gaan..

Het is goed te beseffen dat de hierboven gepresenteerde niveaus van pseudonimisering vooral voor het gemak zijn bedoeld. In feite zijn zij namelijk artificieel, kunstmatig, aangezien het in de praktijk één geheel, een continuüm betreft.

Hieronder volgt een aantal aandachtspunten waar in het pseudonimiserings- en anonimiseringstraject rekening mee gehouden moet worden:

- Verifieer telkens of het pseudonimiseren / anonimiseren goed is verlopen.
  Een eenvoudige controle betreft of er evenveel unieke oorspronkelijke sleutelvelden zijn als geanonimiseerde sleutelvelden, en of beide sleutelvelden evenveel missing values hebben. Een andere controle is de rijen te identificeren die dezelfde oorspronkelijke sleutelwaarde hebben, zodat gekeken kan worden of de geanonimiseerde waarden in die gevallen ook dubbel voorkomen. Dit kan uiteraard alleen als de objecten (personen of organisaties) niet uniek zijn in de dataset.

- Zorg dat je voldoende kennis van de semantiek van de dataset hebt.

  Bepaalde velden kunnen identificerend zijn zonder dat dit op het eerste zicht het geval lijkt. Houd ook rekening met de semantische relatie tussen velden. Het heeft bijvoorbeeld geen

zin om de woonplaatsnamen te vervangen door fictieve steden (een zware vorm van anonimiseren) als de postcode alleen wordt geanonimiseerd door de twee letters ervan af te snijden (een lichtere vorm van anonimiseren/data maskeren). Het is dan nodig de postcode geheel te verwijderen of zwaarder te transformeren.

- Laat iemand anders de gepseudonimiseerde/ geanonimiseerde dataset evalueren.
  De kans dat je identificerende velden over het hoofd ziet of een foutje maakt bij het transformeren is namelijk fors.

- Neem in de analyserapportage de afwegingen en de genomen beslissingen op.
  Op deze manier kun je achteraf aantonen waarom specifieke velden op een bepaalde manier zijn geanonimiseerd.

Bij het opstellen van deze richtlijn is gebruik gemaakt van productionele ervaringen met gegevens uit de Loonaangifte, WGA en BRP. De richtlijn is breed inzetbaar. Echter, er geldt te allen tijde dat er naar de specifieke situatie gekeken moet worden voordat de daadwerkelijke aanpak wordt bepaald.

6. Procesinrichting

Bij het inrichten van processen voor pseudonimiseren / anonimiseren is aantoonbaarheid noodzakelijk. De beveiliging en status van gegevens sets moet te allen tijde aantoonbaar zijn. Van elke pseudonimiserings/anonimiseringsoverweging en -actie wordt een zogenaamd logboek bijgehouden waarin de keuzes als ook argumentatie voor de omgang met de verschillende velden wordt vastgelegd. De Business Security Officer (BSO) zou hierbij betrokken kunnen worden.

De keuze wel/niet pseudonimiseren / anonimiseren wordt bij de ontwikkeling en het testen van software gedeeld met TSC zodat dit wordt meegenomen in het Security Sign Off proces (SSO), waarover door het TSC gerapporteerd wordt, analoog aan de SSD rapportages.

FASE 1 IV domein (Business), (met ondersteuning van TSC)

1. Het proces start bij de business. Hier dient de noodzakelijk kennis van de gegevens en velden in de dataset aanwezig te zijn zodat bepaald kan welke typen velden van belang zijn (precieze, verwijzende, kenmerkende etc.):
   Bepaal aan de hand van het *beoogde datagebruik* of er gepseudonimiseerd / geanonimiseerd moet worden. Functionele tests vereisen doorgaans bijvoorbeeld gefingeerde testgevallen, terwijl voor performance- en stresstests grote hoeveelheden geanonimiseerde productiedata juist voor de hand liggen.
   Analyseer de *aard van de oorspronkelijke data*, waaronder de *koppelbaarheid* en het *identificerend gehalte* van de velden (individueel en in samenhang). Het is van belang hierbij voldoende domeinkennis aan tafel te hebben, aangezien deze analyse kennis vereist van de semantiek van de gegevens en van de desbetreffende bedrijfs-voering.
2. Vervolgens dient te worden bepaald welk niveau (laag, midden, hoog) van anonimiseren nodig is gezien de toepassing van de data buiten de reguliere database (test, analyse of levering aan derde partijen).
   Bepaal voor de dataset als geheel wat het *niveau van pseudonimiseren of volledig anonimiseren* dient te worden en maak beslissingen ten aanzien van de daarbij spelende *trade-offs*. Hierbij is het onder andere van belang hoeveel mensen toegang tot de gepseudonimiseerde / geanonimiseerde dataset krijgen, welke rollen zij bekleden en of het interne danwel externe medewerkers zijn. Ook is het van belang wat de risicoklasse van de gegevens is.
3. Tenslotte dient te worden bepaald welke methodiek wordt toegepast.

(pseudonimisering of anonimisering; zie hiervoor de **beslisboom** (opgenomen in bijlage 3), *deze dient overigens **niet** sec gehanteerd te worden, zonder de overige stappen is de uitkomst uit de beslisboom niet betrouwbaar* !).
Bepaal voor de verschillende velden (variabelen) en records (rijen of observaties) de *wijze van anonimiseren.* Met andere woorden, bepaal de functionele eisen aan de geanonimiseerde set en de optimale mate van 'dataverhaspeling'. Belangrijk hierbij zijn de keuzen t.a.v. zowel *de te anonimiseren velden* als de *anonimiseeroperaties*.

4. Bepaal wie er wordt aangewezen de methodiek (pseudonimisering of anonimisering) toe te passen. Indien dit een functie betreft die in de business ligt, dan draagt de business er de zorg voor dat het bestand in het juiste format bij de juiste instantie wordt aangeleverd.

FASE 2 TSC (met ondersteuning van IV domein, Business)

5. Indien het TSC de instantie is die de methodiek uitvoert (pseudonimisering of anonimisering) en de aangepaste data eventueel plaatst in de betreffende testomgeving, wordt hiervoor een verzoek gedaan via een aanvraag formulier, te verkrijgen op hun site.
6. De transformatie tabel waarmee data gepseudonimiseerd / geanonimiseerd wordt, alsook de data die na anonimiseren beschikbaar komt, mogen pas buiten de extra beveiligde anonimiserings omgeving (alleen toegankelijk voor de betreffende TSC-ers) worden geplaatst, na controle door een security officer (USOC of eigen BSO), door een andere persoon dan diegene die het anonimiseren heeft uitgevoerd maar wel kennis van zaken heeft over het uitgevoerde proces.
7. De betreffende transformatie tabellen moeten t.b.v. anonimisering meteen en bij pseudonimisering na een testfase of indien niet meer nodig worden verwijderd. De bijbehorende logfiles moeten i.v.m. de mogelijkheid voor forensicsh onderzoek minstens 1 jaar bewaard worden.
8. Na anonimiseren dient de gebruikte omgeving te worden geschoond op dusdanige wijze dat de productiegegevens niet meer terug te halen zijn.
9. Het TSC onderhoudt samen met de het IV domein, de business, de aangemaakte testset ten behoeve van toekomstig gebruik/herbruik.

Bijlage 1: Het identificeren van personen

Sommige velden (gegevenselementen of variabelen) identificeren een persoon uniek, terwijl andere velden bij het opzoeken van individuele personen in een dataset meerdere personen tot resultaat kan hebben. Relevant voor het identificerende gehalte van individuele sleutelvelden zijn:

- *Precisie*:
  De scherpte en fijnkorreligheid waarmee het veld naar een unieke persoon kan verwijzen. Dit zegt iets over het onderscheidend vermogen van dit veld. Het BSN is bijvoorbeeld een identificator, omdat het nagenoeg altijd verwijst naar de desbetreffende persoon. Adres is echter ook een scherpe identificator, omdat het een precieze locatie aanduidt van alle mogelijke miljoenen locaties in het land en er meestal maar een zeer beperkt aantal mensen op één adres woonachtig is. Geslacht is minder scherp, aangezien er over het algemeen maar drie keuzes zijn (M, V, O).

- *Beschikbaarheid*:
  Of het veld in veel datasets aanwezig is. Dit verwijst dus naar hoe makkelijk het voor gebruikers van de dataset (zoals testers) is om aan de informatie te komen. Een A-nummer van een buurman of bekende Nederlander is bijvoorbeeld lastig te achterhalen, een adres is redelijk eenvoudig te achterhalen en een geslacht zelfs zeer eenvoudig.

- *Vulling en kwaliteit*:
  Velden die aanwezig zijn in de dataset, maar doorgaans niet of slecht gevuld zijn, zijn minder nuttig voor het identificeren van mensen dan goed gevulde velden.

Verschillende typen velden

Velden die een rol spelen bij het identificeren van personen:

- *Primaire identificerende sleutels*:
  Dit type sleutelveld is belangrijk omdat ze een persoon uniek identificeren en meestal hun bestaansrecht ook aan dit doel ontlenen. Ze zijn doorgaans echter niet bekend bij derden. Voorbeelden zijn: BSN, A-nummer, Onderwijs-nummer, primaire databasesleutels, Rijbewijsnummer, Visumnummer, Bankrekeningnummer, Nummer van mobiele telefoon, Zaak/Case-nummer etc.

- *Secundaire identificerende sleutels*:
  Deze velden zijn belangrijk omdat ze een persoon vaak (bijna) uniek identificeren, vaak in combinatie met een of meer andere velden. Meestal hebben deze gegevens niet primair een identificerend doel. Deze velden zijn regelmatig eenvoudig te achterhalen, en vaak al bekend bij mensen die gegevens van een persoon proberen te achterhalen.

  Een voorbeeld is de combinatie van adres en geboortedatum, waarmee veruit de meeste personen uniek geïdentificeerd kunnen worden. Zelfs als het huisnummer wordt weggelaten, kan de postcode in combinatie met andere velden nog altijd sterk identificerend zijn.

  Andere voorbeelden van secundaire sleutels zijn: Voor- en achternaam, Personeelsnummer, Geboortedatum, Geslacht, Nummer van vaste telefoonlijn, Postcode, Huisnummer (en toevoeging), Gemeente, Locatieomschrijving, e-mailadres, etc. De set van secundaire identificerende attributen die de objecten in een dataset zou kunnen identificeren wordt ook wel *quasi-identifiers* genoemd (Byun, 2007).

- *Hulpvelden*:
  Deze velden zijn op zichzelf niet identificerend, maar kunnen behulpzaam zijn bij de interpretatie van identificerende velden. Voor het goed vergelijken van dynamische identificerende velden is het bijvoorbeeld noodzakelijk de peildatum in ogenschouw te nemen. Om adressen goed te kunnen vergelijken, is het dan ook nodig dezelfde selectiedatum te gebruiken, mogelijk zelfs van meerdere tijdsdimensies. Voorbeelden van dit soort hulpvelden zijn: Datum reële tijd, Datum transactietijd, Kwaliteitskenmerken, etc.

Daarnaast kunnen we nog een ander type veld erkennen:

- *Kenmerkende velden*:
  Dit betreffen attributen van de persoon die vooral beschrijvend zijn, dus iets zeggen over de persoon in plaats van deze te identificeren. Voorbeelden zijn nationaliteit, het al dan niet getrouwd zijn en de CAO-code van iemands inkomstenverhouding. Het interessante is dat ook deze velden tot op zekere hoogte identificerend zijn. Zij hebben op zichzelf vaak geen sterk identificerend potentieel, zelfs niet met een of twee andere kenmerkende velden. Echter, wanneer er sprake is van een grote hoeveelheid kenmerkende velden wordt het steeds onwaarschijnlijker dat de combinatie van waarden veelvuldig in de populatie voorkomt. Daardoor krijgt deze set waarden een steeds groter identificerend potentieel. Bovendien is het mogelijk dat bepaalde waarden van een beschrijvend kenmerk weinig voorkomt (bijvoorbeeld bijzondere klassen of extreme geldbedragen) en daardoor aanknopingspunten bieden voor het identificeren van individuele betrokkenen.

Bijlage 2: Wijze van anonimiseren

Indien een dataset geanonimiseerd dient te worden, dan is de kern van de aanpak:

- De primaire identificerende sleutels dienen altijd geanonimiseerd te worden. Deze sleutels zijn weliswaar niet altijd bekend bij de gebruikers, maar daarvan mag niet uitgegaan worden omdat het doorgaans extern bekende 'betekenisvolle' velden betreffen die juist voor identificatie en koppeling bedoeld zijn.

  Alleen als duidelijk is dat de sleutel niet als zoeksleutel gebruik kan worden, hoeft deze niet geanonimiseerd te worden. Denk aan een situatie waarin maar een handjevol gebruikers de geanonimiseerde data onder ogen zal krijgen, en de primaire sleutels in de buitenwereld niet wijdverbreid of eenvoudig op te zoeken zijn.
  BSN's en A-nummers zijn bijvoorbeeld buiten UWV bekend en moeten om die reden altijd geanonimiseerd worden.

- De secundaire identificerende sleutels dienen geanonimiseerd te worden als ze in de dataset aanwezig zijn in combinatie met de andere velden waardoor ze sterk identificerend worden. Denk hierbij met name aan Postcode in combinatie met Huisnummer, zeker als daarnaast de Geboortedatum en geslacht ook nog zijn opgenomen (hetgeen vaak het geval is). Scherpe 'pseudosleutels' zoals nummers van vaste telefoonlijnen dienen ook geanonimiseerd te worden.

- De hulpvelden hoeven niet geanonimiseerd te worden, aangezien het volstaat om de primaire of secundaire velden ten behoeve waarvan ze gebruikt kunnen worden te anonimiseren.

- De kenmerkende velden (zoals salaris) hoeven niet geanonimiseerd te worden. Echter, indien het Risicoklasse III gegevens betreft, dienen ze voor de zekerheid alsnog o.b.v. vergelijkbare contextwaarde geanonimiseerd worden. Dit is vooral van belang voor de waarden die (in combinatie met andere velden) weinig voorkomen.
  Ook als er veel kenmerkende velden in de set zitten en er reden is aan te nemen dat potentiële gebruikers een match kunnen bewerkstelligen met een bij hen bekende dataset, kan het noodzakelijk zijn een deel van de velden te anonimiseren.

Gegeven een bepaald veld (of combinatie van velden) en afhankelijk van de situatie, kunnen meerdere wijzen van anonimisering gekozen worden:

- *Verwijderen van de waarde*:
  De waarde van het gegeven kan verwijderd worden, bijvoorbeeld door de hele kolom te *verwijderen* uit de dataset, de cellen *leeg* te maken of er een *standaard waarde* in te plaatsen.

  Dit laatste is bijvoorbeeld een mogelijkheid voor performance-testomgevingen. Daarnaast kan *een deel van de celwaarden verwijderd* worden, dus voor maar een deel van de rijen in de dataset. Dit kan met name handig zijn om eenvoudig te identificeren uitbijters of unieke combinaties onherkenbaar te maken. Het verwijderen van gegevens wordt soms ook wel aangeduid als *suppression* (Samarati & Sweeney, 1998). Bedenk wel dat het verwijderen van rijen of kolommen data analyses en testtrajecten danig kan beïnvloeden.

- *Vervangen van de waarde door een willekeurige sleutel*:
  Een sterke manier van het anonimiseren van identificerende sleutels is het vervangen van deze sleutels door een random getrokken nummer. Indien nodig behoudt de datasetbeheerder een overzicht van de relatie tussen de oorspronkelijke sleutel en het nieuwe random nummer.

- *Sleutels in ketens, referentiële sleutel*:
  Als de persoon (of welk object dan ook) in meerdere velden voor kan komen, moet deze uiteraard wel overal dezelfde random sleutel krijgen toebedeeld. In dit geval wordt een primaire identificerende sleutel in het IT-systeem, de applicatieketen of de data analyse gebruikt als *referentiële sleutel*.
  Voor het testen of de analyse betekent dit dat de sleutelfunctie behouden moet blijven nadat het anonimiseringsproces heeft gedraaid. Het anonimiseren dient dan op voorwaarde te gebeuren dat dezelfde random waarde wordt geplaatst op alle locaties waar de sleutel in het systeem als geheel voorkomt. Een voorbeeld hiervan is als een persoon in een rij als

Persoon kan voorkomen en in een andere rij (of dataset) als PartnerVanPersoon. Over het algemeen zal het de bedoeling zijn de relatie tussen de twee personen te behouden, dus ook na het anonimiseren moet dit een stelletje blijven.

- *Transformeren van de waarde*:
  Een bestaande waarde kan gewijzigd worden zodat een fictieve waarde ontstaat. Het is hierbij belangrijk dat er goed wordt nagedacht over de sterkte van de transformatie en of onderlinge relaties behouden moeten blijven.

  Een eerste vorm van het wijzigen van waarden betreft het *willekeurig verplaatsen* van bestaande reële waarden naar de verkeerde entiteiten. Denk aan het plaatsen van een bestaande geldige Postcode-Huisnummer combinatie van een BSN naar een andere BSN. Dit heet soms ook wel *shuffling* (Charles, 2016) of *data swapping* (ICO, 2012).

  Deze aanpak is nuttig omdat het kan garanderen dat de geanonimiseerde dataset reële gegevens blijft bevatten, vooral vanuit het perspectief van technisch testen. Met bepaalde assumpties kan het zelfs de gegevensdistributie en sommige statistische resultaten intact houden (ICO, 2012). Echter, als het om een kleine set gegevens gaat, is het anonimiserend potentieel klein omdat er weinig te 'husselen' valt.

  Gerelateerd hieraan kan men overige elementen, zoals woonplaatsnamen of bedrijfssectoren, voor de zekerheid *verwisselen* met fictieve of buitenlandse categoriewaarden, zodat het voor iedereen absoluut duidelijk is dat het hier een zwaar geanonimiseerde dataset betreft. Denk aan het vervangen van Amsterdam door Gotham City. Dit staat ook wel bekend als *substitution* (Charles, 2016). Deze techniek is niet altijd nodig ten behoeve van anonimiseren, maar kan wel nuttig zijn om de schijn van een datalek te voorkomen.

  Nummervariantie is nog een andere wijze van transformeren, ook wel aangeduid als *variance* (Charles, 2016) of *noise* (ICO, 2012). Hierbij worden numerieke waarden of datums met een willekeurige waarde verhoogd of verlaagd, zodat de oorspronkelijke waarde niet meer scherp te achterhalen valt.
  Een geldbedrag als Loon SV kan bijvoorbeeld binnen een bepaalde bandbreedte gewijzigd worden. En met variantie x en gemiddelde 0 zal deze ruis zich op statistisch geaggregeerd niveau vanzelf opheffen.

  Een vergelijkbare transformatie is de *micro-aggregation* (ICO, 2012), waarbij individuele celwaarden worden vervangen door het gemiddelde van een subset van de rijen.

  Deze vier typen van transformatie zijn niet voor alle soorten gebruik mogelijk: willekeurige verplaatsing is voor technisch testen nuttig, maar niet geschikt voor diverse soorten data analyses. Fictieve waarden houden relaties daarentegen gewoon in stand en kunnen daarom wel voor bepaalde data analyses gebruikt worden, zoals het bouwen van statistische modellen.
  Nummervariantie en micro-aggregation zijn geschikt wanneer gegevens statistisch geaggregeerd worden, maar minder geschikt voor modelbouw omdat je op nominatief niveau minder krachtig associaties kunt identificeren.

  Een volgende vorm van transformeren betreft het *veralgemeniseren* van gegevens, zodat precieze waarden of schaars gevulde klassen niet meer direct of indirect naar individuele objecten kunnen wijzen.
  Denk aan extreme geboortedatums of aan kleine woonplaatsen c.q. bedrijfssectoren, waarvan het al snel duidelijk is wie de oudste inwoners, rijkste personen of grootste bedrijven zijn. Het veralgemeniseren van gegevens wordt soms ook wel aangeduid als *generalization* (Samarati & Sweeney, 1998).
  Denk hierbij aan het wijzigen van een geboortedatum in een geboortejaar, het afkappen van een bankrekeningnummer en het wijzigen van een postcode "2356KL" in "2356".

  In classificaties kan men lang niet altijd direct veralgemeniseren door numerieke of tekstuele waarden eenvoudig af te kappen, maar moeten individuele klassen doelbewust worden samengevoegd.

Een ander voorbeeld van veralgemeniseren is het *discretiseren* van continue variabelen. Dus in plaats van een exact loon in geld zou deze variabele opgedeeld kunnen worden in '0 tot 1500', '1500 tot 4500' en '4500 en hoger'.
Bedenk wel dat er voor het veralgemeniseren van gegevens altijd een prijs betaald wordt. Er gaat namelijk informatie verloren die bij data analyses geëxploiteerd had kunnen worden. Bovendien kan het dataformaat en/of het domein van waarden wijzigen, hetgeen een probleem kan opleveren voor testen.
Dit type transformaties vereist verder dat het veld geen functionele rol als *referentiële sleutel* in het te testen systeem vervult en niet gebruikt wordt in *functionele controles*.

Een laatste transformatietechniek betreft het *hashen* van identificerende gegevenswaarden. Deze versleuteling heeft als voordeel dat je met een eenvoudig aan te roepen transformatie waarden kunt verhullen zonder de onderlinge relaties te verliezen.

Bij voorkeur wordt dan een sterke one-way hash met geheime en sterke salt [2]gebruikt. "Geheim" houdt hier in dat de salt alleen bekend is bij de beheerder van de geanonimiseerde dataset, terwijl "sterk" betekent dat het een lange string (bijvoorbeeld de lengte van de output van de hash) met gewone en speciale karakters betreft. De geheime salt zorgt ervoor dat het voor gebruikers niet meer mogelijk is een voor hen bekend BSN te hashen en het resultaat ervan simpelweg op te zoeken in de geanonimiseerde dataset. Nadelen van hashes zijn dat ze rekentijd in beslag nemen en dat ze het datatype kunnen wijzigen in een vorm waarmee de doelapplicatie in productie wellicht nooit zou werken.

Voor het hashen kan het vanwege praktische doeleinden daarom handig zijn gebruik te maken van format-preserving-hashing. Voor testdoeleinden is dit nuttig omdat het dataformaat behouden blijft en het systeem dus niet vastloopt. Echter, een ander nadeel van hashes is dat ze in theorie te brute-forcen zijn, zeker als het gebruikte hash-algoritme niet erg sterk is. Strikt genomen moeten we bij hashing dan ook spreken van pseudonimisering en niet van anonimisering. En hoe minder sterk het algoritme, hoe korter de hash-string, en hoe groter het risico op een succesvolle hack.

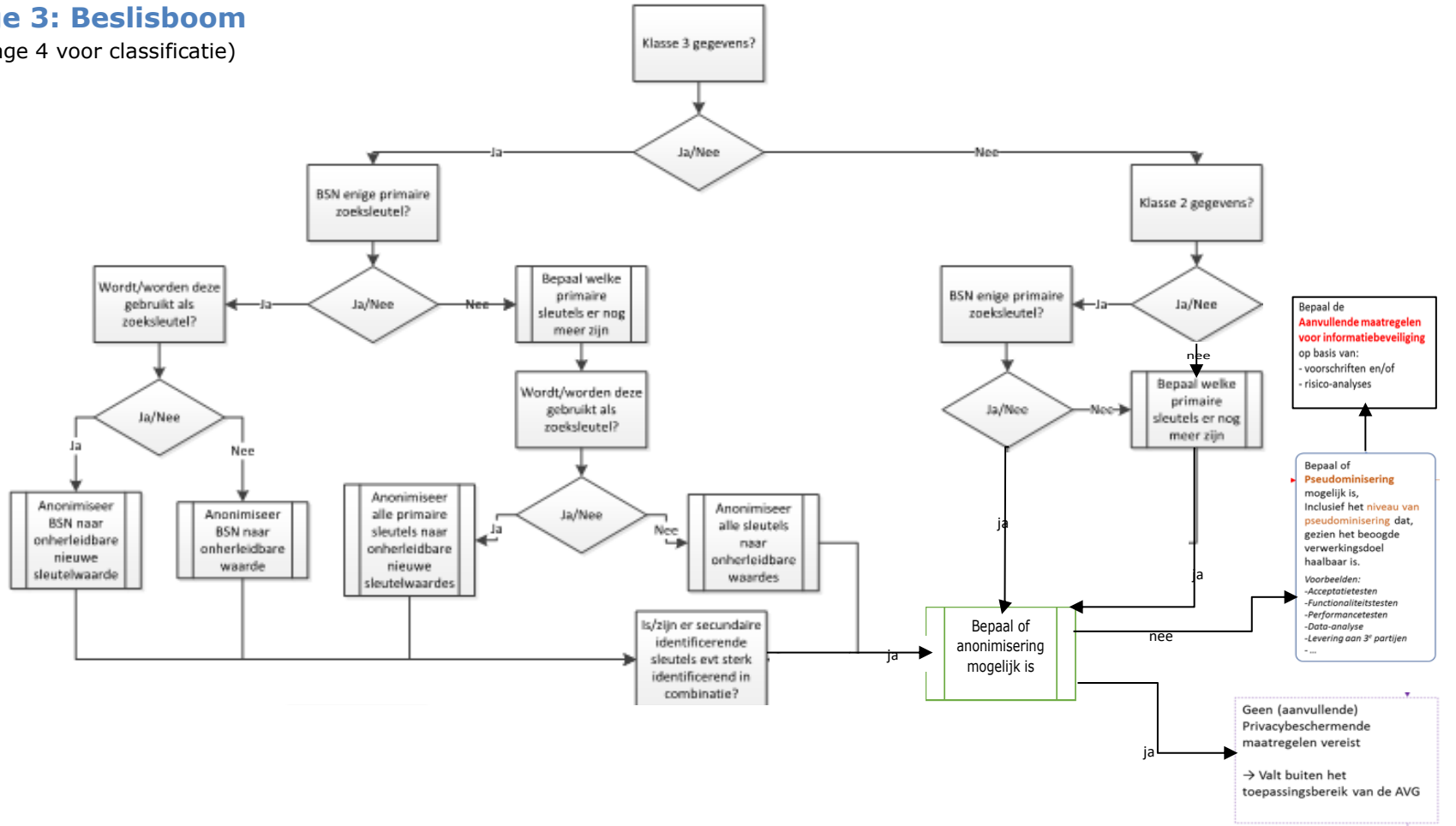
Format-preserving-hashes dienen daarom alleen gebruikt te worden voor velden die niet sterk identificerend zijn. Dit geldt in feite in het algemeen voor zwakke hashes. Zij bieden geen sterke anonimisering, maar hebben wel het voordeel dat ze in een korte string resulteren. Daardoor zijn ze overzichtelijk en nemen ze niet teveel ruimte in beslag.

In verband met de eerder genoemde trade-off zal per situatie bewust gekeken moeten worden welke technieken ingezet dienen te worden. Een en ander hangt af van de desbetreffende dataset (zitten er bijvoorbeeld specifieke additionele gegevens in die ook geanonimiseerd moeten worden?) en het te testen systeem (worden er bijvoorbeeld controles uitgevoerd die veronderstellen dat gegevens niet zijn verhaspeld of zijn losgetrokken van een sleutelwaarde?).

[2] '**salt(ing)'**, oftewel zout(en), is een manier van compliceren, en daardoor extra beveiligen, van het afleiden van sleutels in cryptografie. Salt bestaat uit willekeurige bits die worden toegevoegd aan een wachtwoord (Bron: Wikipedia).

Bijlage 3: Beslisboom

(zie bijlage 4 voor classificatie)



Indien er sprake is van ketenafhankelijkheid zodat de sleutelwaarde(s) in tact moeten blijven en alle mogelijkheden tot pseudonimisering en/of anonimisering zijn beoordeeld waarbij zowel de applicatie/gegevenseigenaar als TSC zijn van mening dat vanwege de keten het BSN of een andere identificerende sleutel intact moet blijven, dan kan een ontheffing worden aangevraagd bij het CISO-Office via de mailbox: ICTSecurityArchitectuur@uwv.nl, waarbij de te nemen beveiligingsmaatregelen die wel mogelijk zijn in ogenschouw worden genomen en waarbij opname in het risicoregister zal plaatsvinden.

Bijlage 4: Classificatie[3]

Vertrouwelijkheidsklassen	
Vertrouwelijkheidsklasse 3 **Strikt vertrouwelijk**	♦ Strikt vertrouwelijke gegevens waarvan ongeautoriseerde onthulling direct of indirect tot ernstige (politieke) schade kan leiden; ♦ Zeer gevoelige gegevens, zoals gegevens waarop een bijzondere of een wettelijk bepaalde geheimhoudingsplicht rust, of gegevens die onder een beroepsgeheim vallen (bijvoorbeeld medisch); ♦ Gegevens van personen naar wie een onderzoek is ingesteld in verband met een mogelijke wetsovertreding.
Vertrouwelijkheidsklasse 2+ **Bijzondere persoonsgegevens**	♦ Bijzondere categorieën van persoonsgegevens zoals genoemd in artikel 9 AVG en de artikelen 22 en 30 wetsvoorstel Uitvoeringswet AVG; ♦ Persoonsgegevens van strafrechtelijke aard zoals genoemd in artikel 10 AVG en de artikelen 32 en 33 wetsvoorstel Uitvoeringswet AVG.
Vertrouwelijkheidsklasse 2 **Vertrouwelijk**	♦ Gevoelige informatie waaronder financieel-economische gegevens in relatie tot de betrokkene en persoonsgegevens, met uitzondering van bijzondere persoonsgegevens zoals genoemd bij Vertrouwelijkheidsklasse 2+; ♦ Vertrouwelijke gegevens die niet voor brede kennisneming zijn bedoeld.
Vertrouwelijkheidsklasse 1 **Basisniveau**	♦ Gegevens die niet zijn bedoeld voor het publiek, maar waarvan ongeautoriseerde onthulling geen schade zal veroorzaken.
Vertrouwelijkheidsklasse 0 **Publiek niveau**	♦ Niet vertrouwelijke gegevens die publiekelijk beschikbaar kunnen worden gesteld; ♦ Openbare persoonsgegevens waarvan algemeen is aanvaard dat deze geen privacyrisico opleveren voor de betrokkene.

[3] Document: UWV BZ BIV classificatie v 1.05

Bijlage 5: Referenties

- Byun, J. (2007). *Toward Privacy-preserving Database Management Systems - Access Control and Data Anonymization*. West Lafayette: Perdue University.
- CBP (2016). *CBP richtsnoeren: beveiliging van persoonsgegevens*. 24 april 2016 opgehaald van URL http://wetten.overheid.nl/BWBR0033572/2013-03-01
- Charles, K. (2016). *Comparing Enterprise Data Anonimization Techniques*. 26 april 2016 opgehaald van URL http://searchsecurity.techtarget.com/tip/Comparing-enterprise-data-anonymization-techniques
- ICO. (2012). *Anonymisation: Managing Data Protection Risk Code of Practice*. Information Commissioner's Office, UK.
- Samarati, P., Sweeney, L. (1998). *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression*. Proceedings of the IEEE Symposium on Research in Security and Privacy (S&P). May 1998, Oakland, CA.
- Sweeney, L. (2000). *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.
- Weber, R.H., Heinrich, U.I. (2012). *Anonymization*. London: Springer.
- Besluit burgerservicenummer
- Besluit kostenvergoeding rechten betrokkene Wbp
- Meldingsbesluit Wbp
- Meldingsregeling Wbp
- Vrijstellingsbesluit Wbp
- Besluit van 5 maart 2012, houdende wijziging van het Besluit kostenvergoeding rechten betrokkene Wbp, het Meldingsbesluit Wbp en het Vrijstellingsbesluit Wbp in verband met vermindering van administratieve lasten.
- de memorie van toelichting bij de wet;
- Wbp-naslag, een naslagwerk van de Autoriteit Persoonsgegevens;
- de Handleiding Wet bescherming persoonsgegevens.

Bijlage 6: Nadere toelichting op pseudonimisering en anonimisering

Encryptie, hashing, pseudonimisering en anonimisering.

Hieronder worden de cryptografische bewerkingen encryptie en hashing toegelicht. Verder wordt aangegeven dat toepassing van deze cryptografische bewerkingen op zichzelf leidt tot pseudonimisering (het vervangen van een identificerend gegeven door een ander identificerend gegeven) en niet tot anonimisering.

Encryptie (versleuteling) en hashing (het omzetten van gegevens in een unieke code) zijn cryptografische bewerkingen die worden toegepast op gegevens met onder meer als doel om deze onbruikbaar te maken voor onbevoegden. Kenmerkend voor encryptie is dat deze bewerking omkeerbaar is: door gebruik van de juiste sleutel kan de oorspronkelijke informatie worden verkregen (decryptie). Encryptie wordt onder meer gebruikt om gegevens te beveiligen bij verzending van gegevens via het internet, bij de opslag van gegevens op draagbare apparatuur en op verwijderbare media zoals USB-sticks en in andere situaties waar gegevens kwetsbaar zijn voor toegang door onbevoegden (bijvoorbeeld gegevens die via het world wide web kunnen worden benaderd). Hashing is een bewerking die van informatie, ongeacht de lengte, een unieke hashcode maakt die altijd even lang is (de lengte is afhankelijk van de gebruikte hashingmethode). Hashing wordt onder meer gebruikt bij de opslag en verwerking van wachtwoorden: op het moment dat de gebruiker een (nieuw) wachtwoord kiest, wordt de bijbehorende hashcode opgeslagen. Wanneer de gebruiker vervolgens inlogt, wordt de hashcode van het ingevoerde wachtwoord vergeleken met de opgeslagen hashcode en krijgt de gebruiker toegang tot het informatiesysteem als de codes overeenkomen.

Cryptografische bewerkingen zijn in principe te 'kraken', wat inhoudt dat onbevoegden toegang kunnen krijgen tot de oorspronkelijke gegevens. Kraken wordt tegengegaan door het gebruik van (combinaties van) de nieuwste cryptografische technieken en door toepassing van zogenoemde salts (extra informatie die bij hashing wordt toegevoegd aan het oorspronkelijke gegeven om het kraken van de hashcode te bemoeilijken). Dit terrein ontwikkelt zich voortdurend en het is zeer goed mogelijk dat een cryptografische bewerking die in de huidige situatie veilig genoeg is dat over een aantal jaren niet meer is. Bij gebruik van cryptografische bewerkingen wordt daarom periodiek beoordeeld of nog steeds wordt voldaan aan de betrouwbaarheidseisen.

Door toepassing van encryptie en hashing kan de mate waarin de verwerkte persoonsgegevens kunnen worden herleid naar "een geïdentificeerde of identificeerbare natuurlijke persoon" worden verminderd. Een voorbeeld is het versleutelen of hashen van klantnummers. Het toepassen van cryptografische bewerkingen op identificerende gegevens leidt op zichzelf tot pseudonimisering (het identificerende gegeven wordt vervangen door een ander identificerend gegeven) en niet tot anonimisering (de gegevens worden omgezet naar "een vorm die identificatie van de betrokkene feitelijk niet langer mogelijk maakt"84).

Nog afgezien van de mogelijkheid dat de gebruikte cryptografische bewerkingen gekraakt worden, leidt het toepassen van cryptografische bewerkingen op identificerende gegevens om de volgende redenen tot pseudonimisering en niet tot anonimisering:

- Encryptie is omkeerbaar. De verantwoordelijke beschikt over de juiste sleutel en is daarmee in staat om decryptie toe te passen op het versleutelde gegeven en daardoor het oorspronkelijke identificerende gegeven te achterhalen.

3 Beveiliging in de Praktijk

- Encryptie en hashing zijn herhaalbaar. De verantwoordelijke beschikt over de oorspronkelijke gegevens en kan door deze opnieuw te versleutelen of hashen de bijbehorende versleutelde gegevens of hashcodes achterhalen.

De verantwoordelijke kan maatregelen treffen om bij gebruik van encryptie of hashing de herleidbaarheid van de betreffende persoonsgegevens te beperken, bijvoorbeeld door de cryptografische bewerking uit te laten voeren door een derde die als enige over de juiste sleutel

beschikt. Van geval tot geval moet een verantwoordelijke vaststellen of dergelijke maatregelen daadwerkelijk leiden tot anonimisering.

Voor anonimisering zijn niet alleen de direct identificerende gegevens van belang. "Het verwijderen van de direct identificerende kenmerken biedt op zichzelf niet altijd voldoende garantie dat geen sprake meer is van persoonsgegevens. Door middel van spontane herkenning, vergelijking van gegevens en/of koppeling aan gegevens uit andere bron, kan immers desondanks, soms zonder bijzonder inspanning, identificatie tot stand worden gebracht." 85 Verder moet bij anonimisering rekening worden gehouden met de stand van de techniek. "Wat [...] bij een bepaalde stand van de techniek als anoniem, want redelijkerwijs niet op een persoon herleidbaar gegeven, kan worden beschouwd, kan door technische ontwikkelingen alsnog een persoonsgegeven worden gelet op de toegenomen mogelijkheden tot herleiding.

Nadere toelichting

Persoonsgegevens kunnen gepseudonimiseerd en geanonimiseerd worden. In het eerste geval is er nog steeds sprake van persoonsgegevens, in het tweede geval niet.

Pseudonimisering
Het doel van pseudonimiseren is het verhullen van iemands identiteit voor derden. Bij pseudonimisering worden identificerende gegevens gescheiden van niet-identificerende gegevens en vervangen door kunstmatige identificatoren. Een voorbeeld van een pseudonimisering is het vervangen van de NAWgegevens van een patiënt in een onderzoeksdatabase door een uniek patiëntnummer. De medische gegevens worden dan gekoppeld aan het patiëntnummer in plaats van aan de NAW-gegevens. Hierdoor is voor buitenstaanders niet zichtbaar wie de persoon is waar de medische gegevens aan toebehoren. Alleen degene die de koppeling kan maken tussen de NAW-gegevens van patiënt en het unieke nummer (bijvoorbeeld de arts) is in staat om de medische gegevens te koppelen aan de geïdentificeerde patiënt.
Gepseudonimiseerde gegevens moeten niet worden verward met anonieme gegevens. Omdat er een koppeling tot stand kan worden gebracht tussen de gepseudonimiseerde gegevens en identificerende gegevens zijn gepseudonimiseerde gegevens onverkort persoonsgegevens. De Verordening is dan ook volledig van toepassing op gepseudonimiseerde gegevens. Wel geeft de Verordening aan dat pseudonimisering een goede maatregel is om persoonsgegevens te beschermen en te beveiligen. Bij het nemen van passende technische en organisatorische maatregelen ter bescherming van persoonsgegevens moet daarom ook pseudonimisering van gegevens overwogen worden.

Anonieme gegevens
De Verordening is niet van toepassing op anonieme gegevens, immers deze gegevens zijn niet terug te voeren op een geïdentificeerde of identificeerbare natuurlijke persoon. Wanneer u persoonsgegevens verwerkt en deze gegevens anonimiseert, dan is de Verordening niet langer van toepassing. Houd er wel rekening mee dat de gegevens daadwerkelijk anoniem zijn en er geen mogelijkheden zijn tot identificatiedoor bijvoorbeeld herleiding, koppeling of deductie. Daarnaast is het anonimiseren van persoonsgegevens zelf wél een verwerkingshandeling.

Het anonimiseren van data is een trade-off (strijd) tussen twee (tegengestelde) ideaaltypen: volledig geanonimiseerde gegevens en volledig representatieve gegevens. Data kun je namelijk dermate extreem verhaspelen dat ze weliswaar compleet onherleidbaar zijn, maar dat testen of data analyses problematisch worden omdat de data niet meer representatief zijn en relaties tussen gegevens verminkt blijken te zijn.