

Veilig omgaan met data: micropseudonimisatie bij NOLAI

Het recente datalek bij een medisch laboratorium laat zien hoe kwetsbaar gevoelige persoonsgegevens zijn. Bij NOLAI worden binnen onderwijsprojecten ook gevoelige gegevens verwerkt. In dit whitepaper leggen we uit hoe NOLAI datalekken voorkomt en hoe dit aansluit bij onze visie op verantwoord en duurzaam datagebruik.



Over de auteur

Job Doesburg is promovendus in het focusgebied Duurzame Data. Zijn promotieonderzoek richt zich op verantwoorde gegevensverwerking in complexe systemen. Onder andere houdt hij zich bezig met PEP-technologie (Polymorfe Encryptie en Pseudonimisatie), een geavanceerde techniek om gegevens te versleutelen en te pseudonimiseren. Hiermee kunnen gegevens veilig worden hergebruikt met minimale privacyrisico's. Goede privacy en security zijn immers essentieel voor verantwoorde toepassing van AI. Eerder schreef Radboud Recharge een [artikel over dit onderzoek](#).

**Job werkt aan dit onderwerp samen met Bernard van Gastel en Erik Poll*

Afgelopen maand werd bekend dat er een groot datalek heeft plaatsgevonden bij een medisch diagnostisch laboratorium dat betrokken is bij het Nederlands bevolkingsonderzoek naar baarmoederhalskanker, waarbij onder andere testuitslagen, BSN's en contactgegevens van zeker 715.000 vrouwen zijn buitgemaakt door criminelen. Er werd in totaal 1.1 miljoen euro losgeld gevraagd om deze data niet openbaar te maken.

Deze hack benadrukt het belang om veilig met dit soort gevoelige gegevens om te gaan. Bij NOLAI's projecten worden ook gevoelige gegevens verwerkt, simpelweg als onderdeel van het onderwijsproces of voor de wetenschappelijke validatie van het prototype. In dit whitepaper geven wij antwoord op de vraag: wat doen wij er bij NOLAI aan om dit soort datalekken te voorkomen, en hoe valt dit samen met ons onderzoek naar verantwoord en duurzaam omgaan met data?

Risico's

Bij NOLAI doen we er alles aan om de kans op een datalek te beperken. Net als elke andere organisatie kunnen we datalekken echter nooit helemaal uitsluiten. Sterker nog, bij NOLAI hebben we een aantal unieke uitdagingen die de kans op datalekken vergroten:

- We hebben een groot aantal projecten waarin gegevens telkens op een andere manier worden verwerkt.
- Bij de gegevensverwerking in deze projecten is een veel en heel diverse partijen betrokken (onderzoekers, scholen, softwareontwikkelaars).
- De ontwikkelsnelheid van deze projecten ligt soms hoog (en om die innovatiekracht niet tegen te gaan willen we ook niet op de rem trappen).

Hoewel er natuurlijk zorgvuldig wordt nagedacht voordat een toepassing wordt gebouwd en gegevens worden verzameld, zorgt dit wel voor uitdagingen. Een medisch laboratorium kan bijvoorbeeld een afgezonderd, beveiligd netwerk aanleggen, waarbij gegevens alleen via dat netwerk beschikbaar zijn. Dat verkleint de kans op datalekken. Maar wanneer een NOLAI-onderzoeker gegevens komt verzamelen in een klaslokaal, is die afhankelijk van het netwerk en de ICT die op die school beschikbaar zijn. Die scholen zitten ook nog eens door het hele land. En wanneer een docent wordt gevraagd een vragenlijst in te vullen, willen we niet dat die docent eerst 5 minuten bezig is met het invoeren van wachtwoorden en beveiligingscodes. Een klaslokaal is nu eenmaal niet een plek waar strenge beveiligingsmaatregelen passend zijn, ondanks dat er veel gevoelige gegevens zijn.

In de informatiebeveiliging hebben we het vaak over risico's. Risico's worden gewogen als kans × impact. Wij proberen zowel de kans op een datalek te verkleinen, als de impact van een eventueel datalek te beperken. Daarmee perken we het risico in. Dit wordt vaak weergegeven in ons vakgebied met de volgende 'formule':

$$\text{risico} = \text{kans} \times \text{impact}$$

Linkbaarheid en identificeerbaarheid

Naast de schaal van een datalek (hoeveel gegevens van hoeveel mensen), is de impact van een datalek vooral afhankelijk van wat voor informatie precies wordt verwerkt. Hoe gevoeliger de gegevens, hoe groter de impact. Het lekken van iemands haarkleur is bijvoorbeeld een stuk minder gevoelig dan hun adres, CITO score, of videobeelden. Ook is het van belang in hoeverre gegevens herleid kunnen worden tot die persoon. Hiervoor zijn een aantal termen belangrijk.

Gegevens zijn linkbaar als ze gekoppeld kunnen worden aan andere gegevens. Als je bijvoorbeeld met een (prepaid) telefoonnummer meerdere personen belt, kunnen die personen door middel van je telefoonnummer informatie bij elkaar leggen, ook al staat nergens van wie dat telefoonnummer is. Hetzelfde geldt voor internetverkeer, door het accepteren van trackingcookies. Wanneer die gegevens uiteindelijk door iemand gekoppeld kunnen worden aan een specifiek natuurlijk persoon, noemen we dat identificeerbaar. Soms kan dat direct, bijvoorbeeld door een naam of e-mailadres, wat als algemeen publiek beschikbare informatie van natuurlijke personen wordt beschouwd. Soms is dat indirect, bijvoorbeeld via het BSN of je IP-adres, waar over het algemeen alleen de overheid en internetprovider over beschikt. Identificeerbaarheid is dus een speciale vorm van linkbaarheid. Door gegevens meerdere keren aan elkaar te linken, kunnen ze uiteindelijk identificeerbaar worden.

Bij een datalek is vaak het grootste probleem niet zozeer de vertrouwelijkheid van de data, maar vooral de identificeerbaarheid. Neem bijvoorbeeld het recente datalek bij het baarmoederhalskankeronderzoek. We stellen voor het gemak dat er drie soorten gegevens zijn buitgemaakt: testuitslagen, e-mailadressen en BSN's.

Ongeacht hoe dit datalek precies tot stand is gekomen, betekent dit dat er ergens binnen een bedrijfsproces van dit lab, waarschijnlijk een tabel heeft bestaan met deze drie kolommen. En dit proces is onvoldoende beveiligd geweest, waardoor de aanvallers ongeautoriseerd toegang hebben kunnen krijgen tot deze tabel.

BSN	E-mail	Testuitslag
111111111	roos.dummy@e-mail.fake	Negatief
123456789	fleur.devoorbeeld@example.com	Negatief
000000000	luna.vandertest@testdomain.nl	Positief
999999999	iris.fictief@mockmail.nl	Negatief

Dataminimalisatie

Stel je voor dat in plaats van deze tabel, enkel de tabel met testuitslagen en BSN's beschikbaar was geweest. Hoewel nog steeds een gevoelig datalek, is dit voor cyber criminelen minder waardevol dan mét e-mailadres, aangezien e-mailadressen publieke informatie zijn en BSN's niet. Phishing of gerichte advertenties op basis van e-mailadres gebeurt continu, op basis van BSN gelukkig niet.

Helaas lekken BSN's soms toch uit, dus beter nog is ook het BSN niet te gebruiken in combinatie met testuitslagen. Dat kan door een tabel te maken met testuitslagen en willekeurige testnummers, en deelnemers zelf bij het afnemen van de test een kaartje met dit testnummer hadden ontvangen. De impact van een datalek van deze tabel was praktisch nihil geweest. We noemen dit testnummer ook wel een pseudoniem: een waarde met beperkte linkbaarheid.

Testnummer	Testuitslag
d9659f22b51e	Negatief
a100b666718	Negatief
b7806828129f	Positief
6254ed63b092	Negatief

Door de dataverwerking dus nét iets anders in te richten, kunnen de privacyrisico's gigantisch worden ingeperkt. Wanneer het BSN en e-mailadres niet verwerkt hoeven te worden, moet dit ook niet gebeuren. Dit idee heet ook wel dataminimalisatie en het ontwerpen van systemen op deze manier noemt men vaak privacy by design. Dit is zelfs een wettelijke eis voor gegevensverwerking volgens de AVG.

Systemeisen en privacy

Het is echter te kort door de bocht om te stellen dat het BSN helemaal niet gebruikt zou moeten worden. In Nederland vinden we het bijvoorbeeld belangrijk dat het laboratorium medische dossier bijhoudt voor iedereen bij wie dit soort medische onderzoeken worden afgenomen. Hiervoor is het opslaan van het BSN zelfs wettelijk verplicht!¹ Daarnaast is het handig dat mensen per e-mail een bericht kunnen ontvangen als hun testuitslag bekend is. Daarvoor is het verwerken van het e-mailadres noodzakelijk.

In de praktijk spelen er dus vaak verschillende tegenstrijdige belangen. Enerzijds móét je bepaalde gegevens verwerken voor een bepaalde functionele eis, anderzijds wil je met oog op privacy die gegevens juist helemaal niet verwerken. Zulke, vaak conflicterende, eisen maken het in de praktijk vaak erg lastig om op een goede manier privacy te waarborgen. Dit merken we ook bij NOLAI. Daarom analyseren we de projecten bij NOLAI nauwkeurig en ontwikkelen we nieuwe technieken om hiermee om te gaan. Een van die technieken is micropseudonimisatie.

Data Field	Filled	Completion %	04D03760D990F2...	5644791A121334...	668AD7A1CD75F...	C423A36C0E417...	64207B1F52DB65...	5030EEFEDA3CD...	1406C3C23BB0B...	EED634FC8F026...	D069EC944BA3...	82CE6562F8E
TeacherInfo_TeacherName	160/160		Present 2025-08-28 16:34:52	Present 2025-03-21 14:25:15	Present 2025-03-25 15:50:11	Present 2025-08-28 16:38:09	Present 2025-08-28 16:28:53	Present 2025-08-28 16:30:56	Present 2025-03-21 14:25:19	Present 2025-08-28 16:35:33	Present 2025-08-28 16:34:16	Present 2025-08-28 16:29
TeacherInfo_EmailAddress	113/160		Present 2025-08-28 16:34:53	Missing	Present 2025-03-25 15:50:12	Missing	Present 2025-08-28 16:28:54	Present 2025-08-28 16:30:57	Present 2025-03-21 14:25:19	Missing	Present 2025-08-28 16:34:17	Present 2025-08-28 16:29
TeacherInfo_ClassName	133/160		Present 2025-08-28 16:34:54	Missing	Missing	Present 2025-08-28 16:38:09	Present 2025-08-28 16:28:56	Present 2025-08-28 16:30:59	Missing	Present 2025-08-28 16:35:36	Present 2025-08-28 16:34:19	Present 2025-08-28 16:29
TeacherInfo_ParticipantNumber	34/160		Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing
Consent_FormData	20/160		Missing	Present 2025-04-03 15:00:59	Missing	Missing	Missing	Missing	Present 2025-03-25 15:28:57	Missing	Missing	Missing
Consent_Bool	55/160		Missing	Present 2025-04-03 15:01:00	Missing	Missing	Missing	Missing	Present 2025-03-25 15:28:58	Missing	Missing	Missing
MinSurvey_Cycle3_BaselineEmailAddress	36/160		Missing	Present 2025-08-28 09:31:36	Missing	Missing	Missing	Missing	Present 2025-08-28 09:31:31	Missing	Missing	Missing
SnapshotLogData_ClassAggregated	45/160		Missing	Present 2025-07-25 14:17:04	Present 2025-07-25 14:17:10	Missing	Missing	Missing	Present 2025-07-25 14:16:50	Missing	Missing	Missing
MinSurvey_Cycle3_BaselineFormData	36/160		Missing	Present 2025-08-29 08:31:37	Missing	Missing	Missing	Missing	Present 2025-08-29 08:31:30	Missing	Missing	Missing
MinSurvey_Cycle4_BaselineFormData	0/160		Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing
MinSurvey_Cycle3_FormData_Week01	24/160		Missing	Present 2025-08-29 08:34:48	Missing	Missing	Missing	Missing	Present 2025-08-29 08:32:28	Missing	Missing	Missing
MinSurvey_Cycle3_OptionalScreenshot_Week01	0/160		Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing
MinSurvey_Cycle3_FormData_Week02	21/160		Missing	Present 2025-08-29 08:34:49	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing
MinSurvey_Cycle3_OptionalScreenshot_Week02	0/160		Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing
MinSurvey_Cycle3_FormData_Week03	25/160		Missing	Present 2025-08-29 08:34:50	Missing	Missing	Missing	Missing	Missing	Missing	Missing	Missing

Figuur 1 Een voorbeeld van hoe we bij een onderzoeksproject data van deelnemers verzamelen met pseudoniemen (kolommen).

Micro-pseudonimisatie

Ook als het opslaan van BSN's en e-mailadressen wettelijk of functioneel noodzakelijk is, betekent dit niet dat de optie met testnummers zinloos is. Zo kunnen testuitslagen, BSN's en e-mailadressen volledig afzonderlijk in drie verschillende tabellen worden opgeslagen (op drie verschillende plekken). De meeste personen of processen binnen het laboratorium zullen waarschijnlijk slechts toegang nodig hebben tot één of twee van deze tabellen, maar niet tot alle drie tegelijkertijd. Wanneer binnen één van die processen dan toch een datalek plaatsvindt, liggen niet direct alle gegevens op straat.

Testnummer	BSN
d9659f22b51e	111111111
a100b666718	123456789
b7806828129f	000000000
6254ed63b092	999999999

Testnummer	E-mail
d9659f22b51e	roos.dummy@e-mail.fake
a100b666718	fleur.devoorbeeld@example.com
b7806828129f	luna.vandertest@testdomain.nl
6254ed63b092	iris.fictief@mockmail.nl

Met deze methode blijven de gegevens in deze drie tabellen alleen nog steeds te linken aan elkaar, bijvoorbeeld als meerdere tabellen uitlekken. We kunnen echter nog een stap verder gaan. In plaats van het gebruiken van een enkel testnummer per deelnemer, gebruiken we voor iedere tabel een ander nummer. Hierdoor worden de gegevens allemaal onlinkbaar aan elkaar. Zelfs als alle drie de tabellen uitlekken, zijn de gegevens praktisch waardeloos (behalve dan dat gezien kan worden dát iemand deelnemer is in het onderzoek).

Deelnemerpseudoniem	Testuitslag
b03e08bc3855	Negatief
e727a94bd55b	Negatief
995fcd328507	Positief
2c9947dd9a90	Negatief

Deelnemerpseudoniem	BSN
4be2367ebaa0	111111111
58a4e0a4b266	123456789
48dcd2e8f3d0	000000000
68dd40521295	999999999

Deelnemerpseudoniem	E-mail
0ce5cff1bce1	roos.dummy@e-mail.fake
02034230e0fe	fleur.devoorbeeld@example.com
8ce2e81b0b61	luna.vandertest@testdomain.nl
ec67f0e9a5aa	iris.fictief@mockmail.nl

Dit lijkt wellicht ridicuul. Want ergens in het systeem moet staan welk e-mailadres en BSN bij welke testuitslag hoort. Moeten we een extra tabel bijhouden waarin we al deze bij-elkaar-horende pseudoniemen opslaan? En wat dan als dié tabel uitlekt?

Gelukkig kunnen we hier vertrouwen op een wiskundig 'trucje' uit de PEP cryptografie² (PEP staat voor polymorfe encryptie en pseudonimisatie). Met dit trucje is het mogelijk om op een veilige manier bijvoorbeeld een deelnemer-pseudoniem voor testuitslagen om te zetten naar een deelnemer-pseudoniem voor e-mailadressen, BSN's, of terug, zonder dat hier een koppeltabel voor nodig is.

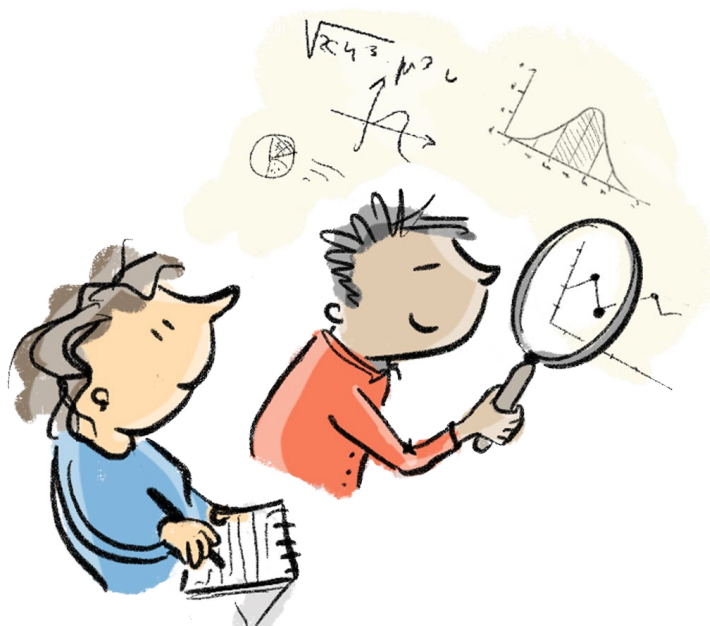


Specifiek is het mogelijk om heel gedetailleerd te bepalen wíé, wannéér en onder welke voorwaarden een specifiek pseudoniem kan omzetten naar een ander pseudoniem, en dit is goed te monitoren en te loggen, en lastig te hacken. Op geen enkel moment bestaat er dus een tabel waarin alle bij-elkaar-horende pseudoniemen staan. Doordat dit ‘trucje’ dus zo goed is te beveiligen en door het omzetten van pseudoniemen alleen toe te staan wanneer het echt noodzakelijk is, kunnen we de impact van datalekken beperken.

(Micro) pseudonimisatie bij NOLAI

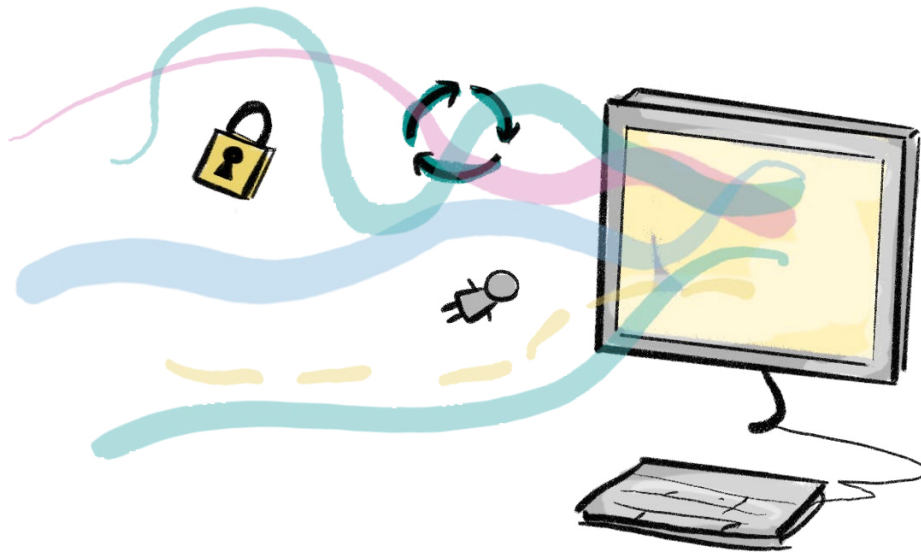
Deze techniek, waarbij gegevens zo veel mogelijk afgezonderd van elkaar worden opgeslagen in verschillende tabellen onder verschillende (micro)pseudoniemen, noemen wij micropseudonimisatie. Dit is een nieuwe techniek die wordt ontwikkeld bij NOLAI, en de implementatie daarvan is open source beschikbaar gesteld.³

Voor sommige onderzoeksprojecten zal van deze techniek gebruik worden gemaakt om onderzoeksgegevens op te slaan. Zo worden gegevens uit verschillende vragenlijsten, interviews, logdata van prototype-applicaties en de toestemmingsformulieren met persoonsgegevens voor deelname in het onderzoek, onder verschillende pseudoniemen opgeslagen. Pas op het moment dat een onderzoeker bepaalde gegevens samen wil analyseren, worden deze omgezet naar een nieuw pseudoniem voor die onderzoeker waarmee de analyse gedaan kan worden. Dit gebeurt alleen voor de gegevens die voor dié analyse nodig zijn.



De toestemmingsformulieren moeten we opslaan zodat we kunnen aantonen dat we alle gegevens rechtmatig opslaan. Deze toestemmingsformulieren bevatten persoonsgegevens, en daardoor zijn deze gegevens identificeerbaar. Daarom wordt toegang tot deze toestemmingsformulieren en persoonsgegevens nooit toegestaan in combinatie met onderzoeksgegevens, zodat de identiteit van deelnemers altijd blijft beschermd. De onderzoeksgegevens zijn dus in principe nooit linkbaar aan de identificeerbare toestemmingsformulieren. Maar zo kunnen we wel aantonen dat we de gegevens rechtmatig opslaan en iemands gegevens ook kunnen verwijderen als daarom wordt gevraagd.

Naast dat deze techniek toegepast kan worden voor veilige dataverwerking voor NOLAI co-creatie projecten en onderzoek, zien we ook toepassingen daarbuiten. Met behulp van micropseudonimisatie kan privacy-vriendelijke opslag en uitwisseling van gegevens tussen organisaties worden gerealiseerd, waardoor de impact van datalekken verkleind wordt. Dit is net zo goed relevant binnen het onderwijs én daarbuiten. Hierdoor draagt micropseudonimisatie in brede zin bij aan het duurzaam omgaan met data.



Duurzaamheid

Een risico van dit soort oplossingen is dat de extra stappen veel extra energie zouden kunnen kosten, waardoor het geen duurzame oplossing is. Daarnaast zou door deze complexiteit wellicht gegevens verloren kunnen gaan. Hier is echter juist tijdens het ontwerp rekening gehouden, zodat het systeem ook sustainable by design is. Want achteraf een systeem duurzamer maken is moeilijker dan vanaf het begin hier rekening mee houden.

Zo zijn er in dit systeem geen kritieke plekken, in vakjargon ook wel single points of failure genoemd, waar de goede werking van het systeem van af hangt. Door systemen redundant (dubbel) uit te voeren blijven de gegevens altijd beschikbaar. Daarnaast is micropseudonimisatie op zo'n manier ontworpen dat de extra stappen minimaal extra energie kosten. We hebben dit onder andere doorgemeten in het Software Energy Lab van de Radboud Universiteit en concluderen dat het extra energieverbruik zeer beperkt en acceptabel is. Daarmee beschouwen we deze techniek ook als duurzaam: met minimale extra energie zijn de gegevens beter beveiligd en blijven ze goed beschikbaar.

Conclusie

De kans op datalekken kan nooit volledig worden uitgesloten. Daarom ontwikkelen we bij NOLAI nieuwe technologie waarmee de impact van datalekken kan worden beperkt. Door toepassing van (micro)pseudonimisatie zijn onderzoeksgegevens in ieder geval nooit identificeerbaar of linkbaar, waardoor de gevolgen van een datalek worden geminimaliseerd. Deze technologie van micropseudonimisatie is niet alleen waardevol voor ons onderzoek bij NOLAI, maar zeker ook daarbuiten.

Bronnen

1. Lees ook <https://blog.xot.nl/2025/08/26/de-hack-bij-clinical-labs-had-eenvoudig-voorkomen-kunnen-worden/index.html> voor een interessante observatie hoe de wet privacy hier in de weg staat, en hoe *privacy by design* ook van toepassing is op het ontwerp van *wet- en regelgeving*.
2. Verheul, E.R. en Jacobs, B., Polymorphic encryption and pseudonymisation in identity management and medical research, In Nieuw Archief voor Wiskunde (NAW 5/18 nr. 3 september 2017). <https://www.nieuwarchief.nl/serie5/pdf/naw5-2017-18-3-168.pdf>
3. Zie onder andere <https://github.com/NOLAI/libpep>