DATA ARCHIPEL – ANALYSE OP GEKRUISTE PERSOONSGEGEVENS

KRISTOF VERSLYPE

Abstract. Analytics op persoonsgegevens moet steeds in overeenstemming gebeuren met de privacywetgeving, waarbij principes zoals finaliteit, proportionaliteit, transparantie en information security practices gerespecteerd dienen te worden. De fragmentatie van de overheid levert een bijkomende uitdaging op. Wanneer namelijk persoonsgegevens uit verschillende bronnen gekruist worden in één centraal datawarehouse, blijven de aanleverende overheidsbedrijven enerzijds wel verantwoordelijk voor de persoonsgegevens, maar anderzijds verliezen ze alle controle erop.

Deze tekst presenteert op een toegankelijke manier het door Smals Research ontwikkelde concept van de *data archipel*, dat gebruikmakend van cryptografie drie moeilijk te combineren zaken realiseert: 1) het kruisen van persoonsgegevens voor analyse is vlot mogelijk, 2) de privacy van de betrokken burger wordt beschermd en 3) de aanleverende overheidsbedrijven behouden controle op de persoonsgegevens waar ze verantwoordelijk voor blijven.

Résumé. L'analytique de données personnelles doit toujours s'effectuer en conformité avec la législation de la vie privée, qui impose le respect de principes tels que la finalité, la proportionnalité et la transparence mais aussi des pratiques de sécurisation de l'information. La fragmentation de l'État entraîne un défi supplémentaire. En effet, lorsque des données personnelles issues de sources diverses sont croisées dans un datawarehouse central, les entreprises publiques qui livrent des données personnelles demeurent responsables de ces données, mais elles en perdent totalement le contrôle.

Ce texte présente en termes clairs le concept de data archipel, que l'équipe Smals Research a développé et qui, au moyen de la cryptographie, permet d'atteindre trois objectifs difficilement combinables : 1) le croisement de données personnelles à des fins d'analyse est facilement réalisable, 2) la confidentialité du citoyen concerné est protégée et 3) les entreprises publiques gardent le contrôle des données qu'elles livrent et dont elles sont responsables.

Contents

1. Introductie

Een onderzoeksteam aan een universiteit wil toegang tot een set (niet-geaggregeerde) persoonsgegevens van alle getrouwde burgers met een jaarlijks inkomen van minstens € 50 000 en die in 1990 of daarna geboren zijn. Het team wil meer bepaald over elk van deze burgers een aantal medische, financiële en demografische persoonsgegevens. De gevraagde gegevens worden echter beheerd door verschillende overheidsinstellingen. De logische oplossing om aan dergelijke vragen tegemoet te komen is dan ook een centraal datawarehouse waar de betrokken overheidsinstellingen de gevraagde persoonsgegevens die ze beheren aan bezorgen. Deze overheidsinstellingen zijn begrijpelijkerwijs terughoudend om de controle op te geven over persoonsgegevens waar ze wettelijk verantwoordelijk voor blijven. De centrale vraag in deze tekst is dan ook: "*Hoe kunnen we persoonsgegevens afkomstig van verschillende overheidsinstellingen op een efficiënte, zo veilig mogelijke manier kruisen, waarbij de privacywetgeving gerespecteerd wordt en de aanleverende overheidsorganisaties controle blijven behouden over de persoonsgegevens die ze beheren?*"

Eén van de voornaamste evoluties in het huidige digitale landschap is de toenemende collectie en verwerking van gegevens, waaronder ook vaak persoonsgegevens. De potentiële waarde die uit deze data geëxtraheerd kan worden is enorm, niet enkel voor privé- maar ook voor overheidsbedrijven. Analyse van deze gegevens kan dus een essentiële rol spelen in het vervullen van de taken van deze laatste. Voorbeelden zijn het verbeteren van de gezondheid van de burgers, het bestrijden van fraude en ondersteuning bij beleidsbeslissingen. Ook voor onderzoekers is dergelijke data van onschatbare waarde.

Maar de verwerking van gegevens is natuurlijk enkel toegelaten indien het in overeenstemming is met de van toepassing zijnde wetgeving. In het geval van persoonsgegevens is dit voornamelijk (maar niet steeds uitsluitend) de privacywetgeving gebaseerd op de Europese richtlijn EC 95/46/EC [B92][1]. Deze richtlijn definieert een aantal principes. Zo is het niet toegestaan om persoonsgegevens te verwerken op een manier die onverzoenbaar is met de redenen waarvoor ze verzameld werden (*finaliteit*). Ook mag er niet meer data verzameld worden dan strikt noodzakelijk voor de doelstelling waarvoor de data verzameld werd (*proportionaliteit*). Verder heeft de burger (data subject) het recht te weten welke verwerkingen er plaatsgevonden hebben op persoonsgegevens die op hem van toepassing zijn (*transparantie*). Ook zal de eigenaar van de data, de data controller, aansprakelijk gesteld worden in geval van een data breach door nalatigheid (*information security practices*). Dat dit laatste geen overbodige eis is, blijkt uit de talrijke incidenten waarbij overheidsdata gelekt is. Het bekendste recente voorbeeld hiervan is de OPM hack [M15].

Binnen een overheidscontext zijn er twee eigenschappen die de verwerking van gegevens in overeenstemming met de wetgeving extra uitdagend maken. Ten eerste werken overheden zeer vaak met persoonsgegevens. Ten tweede kent de overheid een grote mate van fragmentering: elk overheidsbedrijf is en blijft verantwoordelijk voor een specifieke set van persoonsgegevens. Juridisch is een overheidsbedrijf dus de *data controller* van deze persoonsgegevens.

Het combineren van persoonsgegevens afkomstig van verschillende overheidsbedrijven in één groot traditioneel datawarehouse heeft twee grote

[1] We verwachten binnenkort de nieuwe Europese privacyverordening. De verwachting is niet dat de in deze tekst vermelde principes afgezwakt zullen worden.

nadelen. Ten eerste verliest het aanleverende overheidsbedrijf (data controller) alle controle over de persoonsgegevens waarvoor het juridisch verantwoordelijk is en waarvan het niet wil dat het tegen zijn belangen in gebruikt wordt. Ten tweede kan een data breach catastrofale gevolgen hebben voor de privacy van de betrokken burgers en voor de reputatie van de betrokken overheidsbedrijven[2].

De terughoudendheid van overheidsbedrijven om hun data te kruisen in één groot datawarehouse is dus terecht. Vanuit een juridisch standpunt en wat de veiligheid betreft is een klassiek datawarehouse dus een '*digital data dystopia'* voor zowel het aanleverende overheidsbedrijf als de burger. Tegelijkertijd is er wel een behoefte om persoonsgegevens afkomstig van verschillende bronnen te kruisen voor geavanceerde gegevensanalyse, maar dan wel binnen het juridisch kader.

Dit brengt ons tot de uitdaging die in deze tekst behandeld wordt: het toelaten van geavanceerde analyse van persoonsgegevens afkomstig van verschillende bronnen, met respect voor de privacy, terwijl 1) de aanleverende (overheids)bedrijven de controle over data waarvoor ze verantwoordelijk zijn behouden en 2) de impact in geval van een data breach geminimaliseerd wordt. Tegelijkertijd biedt de voorgestelde oplossing, de *data archipel*, een hoge graad van flexibiliteit, bijvoorbeeld bij sleutelbeheer en deanonimisatie, zodat ze in een brede waaier aan contexten ingezet kan worden.

Deze tekst tracht op een toegankelijke manier de principes van de data archipel uit te leggen. We verwijzen naar het technisch *rapport Data Archipelago - Reconciling privacy and analytics on multi-source PII* [VD16] voor een formele en diepgaande technische beschrijving. Dit technisch rapport is tot stand gekomen in nauwe samenwerking met het departement Computerwetenschappen van de KU Leuven.

Sectie 2 bespreekt de beperkingen van anonimisatietechnieken, sectie 3 gaat dieper in op de basisconcepten van de data archipel en schetst hoe dit gerealiseerd kan worden, sectie 4 bespreekt een aantal uitbreidingen, sectie 5 toont de proof-of-concept en sectie 6 somt enkele beperkingen en uitdagingen op. Sectie 7, ten slotte, geeft de conclusies.

2. De illusie van anonimisatie

Data-anonimisatietechnieken zoals *k-anonymity*, *l-diversity* en *t-closeness* werden naar voor geschoven als de manier om analyse van persoonsgegevens met privacy te verzoenen. Het heeft tot doel om onomkeerbaar de link tussen de gegevens en het individu te verwijderen. Daartoe worden 1) de directe identifiers, zoals sociale zekerheidsnummers, verwijderd of vervangen en 2) pseudo-identifiers verwijderd of veralgemeend, of wordt er ruis aan toegevoegd. Pseudo-identifiers zijn combinaties van attributen die samen vaak maar tot één persoon te herleiden zijn. Een typisch voorbeeld is de combinatie geslacht, geboortedatum en postcode. Een voorbeeld van vervagen van persoonsgegevens is het omzetten van een geboortedatum (vb. 03/04/1967) naar een leeftijdscategorie (vb. tussen 40 en 50 jaar). Eens gegevens wettelijk geanonimiseerd zijn, worden ze niet meer beschouwd als persoonsgegevens, is de privacywetgeving dus niet meer van toepassing en wordt de schade sterk beperkt in geval van een data breach.

[2] In de huidige manier van werken is de situatie zelfs nog moeilijker te overzien en te beheren, daar in de praktijk niet één groot (gedeeld) datawarehouse wordt opgezet, maar één per instelling (die instellingsdoeleinden bereikt d.m.v. kruising van eigen data met data afkomstig van andere instellingen).

Helaas resulteert het toepassen van anonimisatietechnieken op hoogdimensionale gegevens[3] niet tot gegevens die tegelijkertijd nog bruikbaar en juridisch geanonimiseerd zijn. Zelfs als de gegevens sterk aan waarde verloren hebben door het toepassen van de anonimisatiechnieken, kunnen ze nog steeds linkbaar zijn aan één individu [M13]. Het Witte Huis bv. beschouwt dan ook terecht anonimisatietechnieken als verouderd in het tijdperk van big data, hoewel ze in het verleden wel nuttig waren [PCAST14].

Bovendien is het zeer moeilijk om na te gaan of een dataset voldoende geanonimiseerd is, wat geïllustreerd wordt d.m.v. twee voorbeelden.

1) AOL publiceerde in 2006 20 miljoen 'geanonimiseerde' zoekopdrachten van 650 000 gebruikers. Al gauw werd de identiteit van meerdere gebruikers onthuld [BT06].
2) In 2007 publiceerde NetFlix 'geanonimiseerde' filmbeoordelingen van 500 000 gebruikers. Door deze gegevens te kruisen met publieke data op IMDb konden records gedeanonimiseerd worden [NS09].

Gegeven het gebrek aan informatie over de kennis die de aanvaller heeft, zowel in termen van algoritmes als in termen van data waar deze over beschikt, lijkt een formele analyse om de anonimiteit te meten onmogelijk. Paul Ohms stelde dan ook: "*De robuuste anonimisatieassumptie is niet fundamenteel incorrect, maar wel erg gebrekkig (deeply flawed)*" [O09].

Samengevat zijn anonimisatietechnieken onvoldoende in onze context. In een uitgebreider artikel "*Big data & krakend ijs onder anonimisatie*" [V15] wordt dieper ingegaan op deze problematiek.

Het gevolg is dat het zogenaamde "*small-cell"*-probleem onoplosbaar is in een big data context. Small cells ontstaan wanneer een dataset met persoonsgegevens, maar zonder identifiers (identificatiesleutels) zoals naam en INSZ-nummer, aan een externe partij, zoals een onderzoeker, gegeven wordt en de individuele records in de dataset veelal maar van toepassing zijn op één persoon. Mits extra informatie is het zo mogelijk om het record te linken aan een geïdentificeerd individu.

> Stel bijvoorbeeld dat een onderzoeker beschikt over een set records; één record per betrokken burger. Elk record bevat onder meer de wijk, geboortejaar, haarkleur en bloedgroep en extra medische gegevens. Het is waarschijnlijk dat de combinatie wijk-haarkleur-geboortejaar voor bepaalde records maar kan betrekking hebben op één persoon. Indien we zo een persoon kunnen identificeren, bijvoorbeeld omdat we hem/haar persoonlijk kennen, kunnen we ook de extra medische gegevens aan deze persoon linken.

[3] In tegenstelling tot ééndimensionale data zoals salaris, geboortedatum en lengte zijn hoogdimensionale data enkel uit te drukken met een groot – vaak stijgend – aantal getallen. Een typisch voorbeeld is location tracking; het bijhouden van de locatie van individuen op verschillende tijdstippen. Elke meting bestaat uit twee dimensies (tijd + locatie) en bij één meting per uur – wat zeer conservatief is – zitten we al aan 48 dimensies per dag.

3. Data Archipel

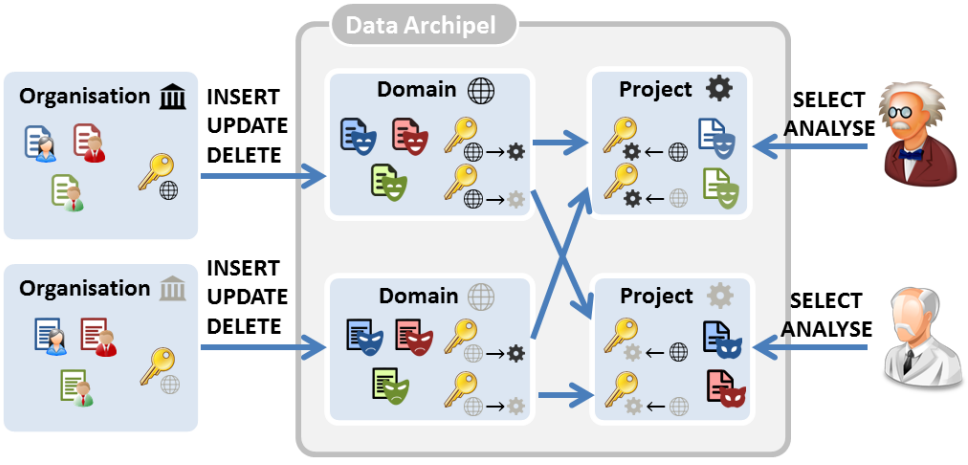
Deze sectie bespreekt de data archipel, een concept dat wij voorstellen als antwoord op bovenstaande problematiek. Eerst geven we een overzicht op hoog niveau, daarna wordt een belangrijke bouwsteen, de pseudoniemconversiefunctie beproken en wordt geschetst hoe dit gerealiseerd kan worden m.b.v. elliptische krommen.

3.1. Overzicht

De data archipel wordt weergegeven in figuur 1. De twee types deelnemers zijn organisaties (links) en onderzoekers (rechts). Een *organisatie* levert aan de data archipel dat deel van de persoonsgegevens die ze beheert dat in aanmerking komt om in specifieke analyseprojecten gebruikt te worden. Een *onderzoeker* wil persoonsgegevens afkomstig van verschillende organisaties kruisen om ze vervolgens te analyseren.

De data archipel zelf bestaat uit geïsoleerde *eilanden* die ofwel een *domein* (links) zijn, ofwel een *project* (rechts). Een domein slaat enkel persoonsgegevens op die onder de verantwoordelijkheid vallen van één specifieke organisatie. Een domein is relatief permanent en heeft lage performantievereisten. Een project is de omgeving waar analyse gebeurt op gekruiste persoonsgegevens afkomstig van verschillende domeinen. Een project bevat enkel de minimaal noodzakelijke data. Projecten zijn tijdelijk en worden zo gauw als mogelijk vernietigd (of gearchiveerd indien wettelijk verplicht). Een project heeft hoge performantievereisten voor complexe data-analyse. Toegang tot projecten door onderzoekers is beperkt en wordt gemonitord om maximaal de persoonsgegevens te beschermen.

Stel, bijvoorbeeld, dat uw jaarlijks inkomen beheerd wordt door de ene overheidsorganisatie en uw burgerlijke staat door de andere. Uw jaarlijks inkomen komt in het ene domein terecht, terwijl uw burgelijke staat terechtkomt in het andere domein. Een project kan uw burgerlijke stand bevatten alsook een inkomenscategorie, wat eenvoudig afgeleid kan worden uit uw jaarlijks inkomen. Het project kruist dus data afkomstig uit minstens twee domeinen.



Figuur 1. Overzicht van de data archipel

De isolatie tussen de verschillende eilanden wordt gemaximaliseerd op zowel hardwareniveau d.m.v. (bestaande) technieken zoals containerization als op het

niveau van de persoonsgegevens door het minimaliseren van linkbaarheden tussen individuele records op verschillende eilanden. Dergelijke linkbaarheden onstaan wanneer verschillende eilanden voor dezelfde burger één of meer van de volgende zaken delen:

1) een *identifier* zoals een INSZ-nummer, dat uniek een persoon identificeert,
2) een *pseudoniem*, dat toelaat om verschillende records van dezelfde burger aan elkaar te linken, maar niet aan de fysieke persoon zelf
3) een *pseudo-identifier* die, zoals eerder reeds beschreven, een combinatie van attributen is, die vaak maar tot één persoon te herleiden is.

Anders gezegd, uw record in het ene domein dat o.a. uw jaarlijks inkomen bevat mag niet linkbaar zijn aan uw record in het andere domein dat o.a. uw burgerlijke status bevat. Ook moet het erg moeilijk gemaakt worden om uw record in een project te linken aan uw records in de verschillende domeinen of aan records van u in andere projecten.

Identifiers (zoals het INSZ-nummer) worden niet toegelaten in de data archipel. Dit kan afgedwongen worden d.m.v. reguliere expressies die identifiers detecteren in records die naar de data archipel gezonden worden. Vervolgens kan het record waartoe de identifier behoort geblokkeerd worden.

Duidelijke afspraken tussen organisaties vermijden datareplicatie over meerdere domeinen. Dit resulteert in attribuutpartities. Zo kunnen de domeinen van authentieke bronnen zoals het rijksregister exclusief attribuutwaarden (of afgeleiden) aanleveren aan projecten. Als het rijksregister in haar domein dus geboortedatum en woonplaats bijhoudt, zal geen enkel ander domein deze of afgeleide attributen bevatten.

Wel zullen hoogstwaarschijnlijk attribuutlinkbaarheden ontstaan tussen projecten onderling en tussen projecten en domeinen. Dit is onvermijdelijk. Het is dan ook – net zoals in reguliere analyseprojecten – vereist dat projecten goed beschermd worden. Maar in tegenstelling tot reguliere analyticsprojecten bevatten projecten in de data archipel gelukkig enkel de strikt noodzakelijke data en zijn ze tijdelijk.

Naarmate meer attributen in een domein aanwezig zijn, groeit de impact in het geval van een data breach. Een domein dat op een gegeven moment te veel attributen bevat kan dan ook gesplitst worden in meerdere gescheiden domeinen die elk afzonderlijk verantwoordelijk zijn voor een deel van de attributen. Deze nieuwe domeinen zijn dus onderling onlinkbaar.

Deze subsectie schetste op hoog niveau de data archipel. De rest van deze tekst gaat in op het garanderen van onlinkbaarheden van records over dezelfde burger tussen eilanden op basis van pseudoniemen d.m.v. cryptografie. Het kruisen van gegevens uit meerdere domeinen voor een project is mogelijk op een efficiënte manier, maar vereist de geheime sleutels – en dus de medewerking van – de betrokken domeinen.

3.2. Pseudoniemconversiefunctie

Deze subsectie gaat in op de pseudoniemconversiefunctie, wat een essentiële component is van de data archipel. Zoals geïllustreerd in figuur 1 heeft elke organisatie (overheidsbedrijf) een sleutel, heeft elk domein één sleutel per project waaraan het data levert en heeft een project één sleutel per aanleverend domein. Samen met een identifier of een pseudoniem wordt een dergelijke sleutel als input

gegeven aan de pseudoniemconversiefunctie f. De output is een nieuw pseudoniem. We krijgen dus:

$$pseudonym \leftarrow f(sleutel, identifier) \text{ en}$$
$$pseudonym \leftarrow f(sleutel, pseudonym).$$

Er wordt een onderscheid gemaakt tussen (permanente) domeinpseudoniemen, (tijdelijke) transferpseudoniemen en (finale) projectpseudoniemen. Domeinpseudoniemen zijn door organisaties gegenereerd op basis van een identifier en worden bewaard in een domein. Transferpseudoniemen zijn om praktische redenen niet weergegeven in figuur 1, maar worden gegenereerd door domeinen op basis van domeinpseudoniemen. Een project ontvangt transferpseudoniemen die het converteert naar projectpseudoniemen. We krijgen nu:

$$domain\_pseudonym \leftarrow f(sleutel, identifier),$$
$$transfer\_pseudonym \leftarrow f(sleutel, domain\_pseudonym) \text{ en}$$
$$project\_pseudonym \leftarrow f(sleutel, transfer\_pseudonym)$$

Twee domeinpseudoniemen afkomstig van dezelfde burger (maar in verschillend domein) zullen verschillend zijn, maar de daaruit afgeleide projectpseudoniemen zijn identiek binnen eenzelfde project, **waardoor gegevens over dezelfde persoon niet linkbaar zijn op het niveau van de domeinen, maar gelinkt kunnen worden op het niveau van een specifiek project.**

De functie f moet voldoen aan verschillende eigenschappen:

- *Onlinkbaarheid*. Verschillende pseudoniemen die afkomstig zijn van dezelfde identifier zijn onlinkbaar aan elkaar en aan de oorspronkelijke identifier.
- *Lokale uniekheid*. Pseudoniemconversies op het niveau van de projecten toegepast op onlinkbare transferpseudoniemen (komende van verschillende domeinen) resulteren in hetzelfde projectpseudoniem als en slechts als de transferpseudoniemen afgeleid zijn van dezelfde oorspronkelijke identifier.
- *Botsbestendig*. Het toepassen van f met dezelfde sleutel op twee verschillende pseudoniemen of identifiers resulteert steeds in verschillende pseudoniemen.
- *Deterministisch*. Het meermaals toepassen van f op dezelfde input resulteert steeds in dezelfde output.
- *Inverteerbaar*. Gegeven de sleutel en de output-waarde van f is het gemakkelijk om daaruit weer het oorspronkelijke pseudoniem of de oorspronkelijke identifier die als input gegeven werd, af te leiden.

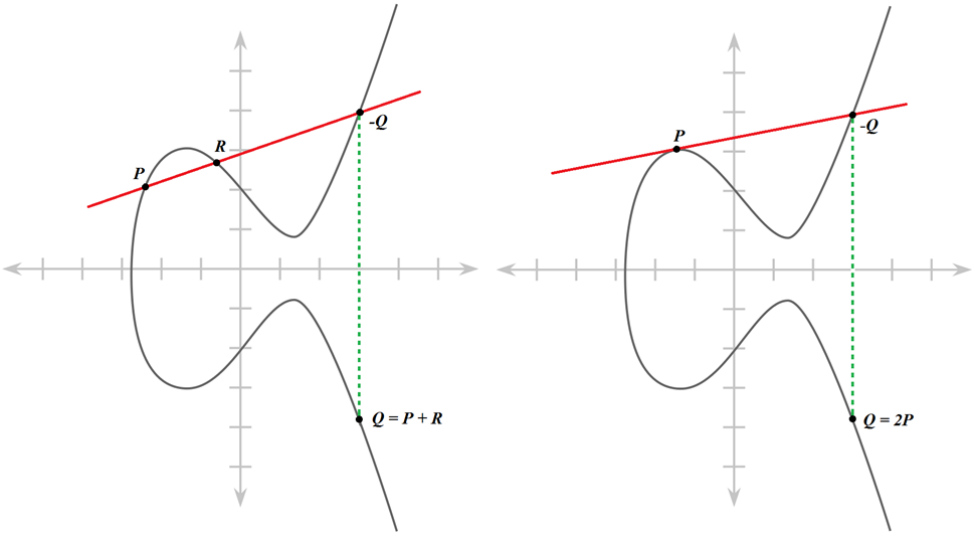
We weten nu welke eigenschappen een pseudoniemconversiefunctie moet hebben, maar we weten nog niet hoe we een dergelijke functie kunnen realiseren en gebruiken in de data archipel. Dit wordt in de twee volgende subsecties toegelicht.

3.3. Elliptische krommen

De pseudoniemconversiefunctie kan gerealiseerd worden m.b.v. elliptische krommen, wat wiskundige structuren zijn die vaak in hedendaagse cryptografie gebruikt worden [M15a, M15b]. In deze subsectie geven we de essentie van elliptische krommen. Vervolgens geven we in de volgende subsectie intuïtief mee hoe deze elliptische krommen de pseudoniemconversiefunctie kunnen realiseren.

Beide subsecties kunnen overgeslaan worden door lezers die minder geïnteresseerd zijn in de wiskunde achter het concept van de data archipel.

Elliptische krommen en de operatie die we erop definiëren zijn geïllustreerd in figuur 2. Op punten P en R op een elliptische kromme kan een operatie '+' gedefinieerd worden, wat resulteert in het punt Q ($P + R = Q$) als volgt: we trekken de lijn die punten P en R snijdt. Deze lijn zal de curve opnieuw kruisen in een punt $-Q$. We spiegelen dit punt over de x-as wat resulteert in het punt Q. Indien we de '+' operatie toepassen op hetzelfde punt P ($P + P = Q$), nemen we de raaklijn op het punt P, die de curve zal snijden in het punt $-Q$. Dit punt spiegelen we opnieuw over de x-as, wat resulteert in Q.

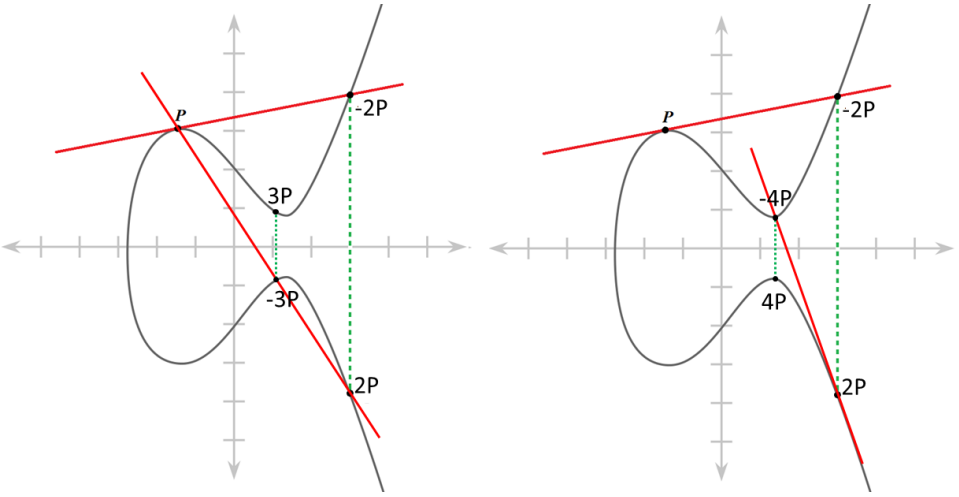


Figuur 2. Elliptische krommen in het reële vlak en de berekening van $P + R = Q$ (links) en $P + P = Q$ (rechts)

Hoe kunnen we nu $3 \cdot P = P + P + P$ berekenen, gegeven een punt P? Dit is geïllustreerd links in figuur 3. Eerst berekenen we net zoals daarnet het punt $2P = 2 \cdot P = P + P$. Vervolgens gaan we punten P en $2P$ met elkaar optellen door opnieuw de rechte die deze twee punten snijdt te bepalen. Deze rechte zal opnieuw de elliptische kromme snijden in een derde punt $-3P$. Dit punt spiegelen we opnieuw over de x-as, wat resulteert in het punt $3P = 3 \cdot P = P + P + P$.

Op een gelijkaardige manier kan ook $4 \cdot P = P + P + P + P$ berekend worden. Dit is geïllustreerd rechts in figuur 3. We berekenen opnieuw eerst het punt $2P = 2 \cdot P = P + P$. We trekken de raaklijn door dit punt $2P$, wat de kromme zal snijden in het punt $-4P$. We spiegelen dit punt over de x-as en bekomen aldus $4P = 4 \cdot P = P + P + P + P$.

We kunnen zo verder gaan en $n \cdot P = Q$ berekenen voor willekeurig grote n. De centrale aanname in cryptografie met elliptische krommen is dat, hoewel $n \cdot P = Q$ erg efficiënt te berekenen is, **er geen enkele haalbare manier bestaat om gegeven P en Q de waarde n te vinden voor voldoende grote waarden van n (vb. 256 bits).** Cryptografie met elliptische krommen wordt algemeen gezien als de opvolger van cryptografie gebaseerd op RSA aangezien het veel efficiënter is. Of anders gezegd, het kan hetzelfde niveau van veiligheid garanderen met veel kortere sleutels.



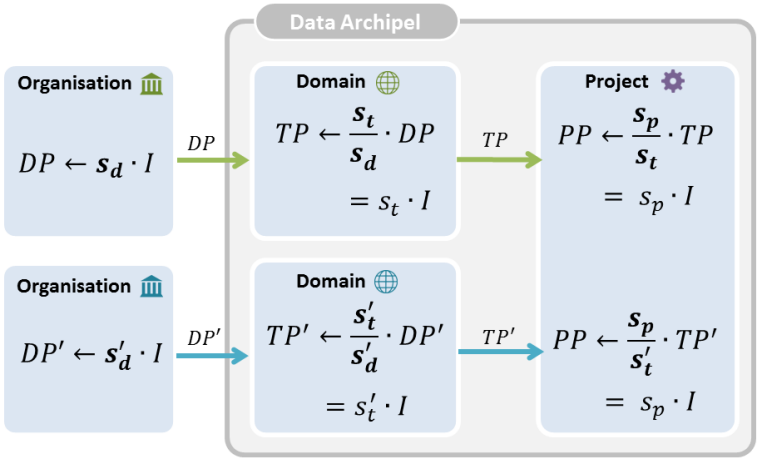
Figuur 3: De berekening van $3 \cdot P = P + P + P$ (links) en $4 \cdot P = P + P + P + P$ (rechts) gegeven een punt P op een elliptische kromme.

3.4. Realisatie

In deze subsectie tonen we intuïtief hoe de pseudoniemconversiefunctie gerealiseerd kan worden d.m.v. elliptische krommen. In figuur 4 zijn de sleutels in het vet aangegeven. Bemerk daarbij dat hoewel deze meestal uit een breuk bestaan, de eigenaar de afzonderlijke geheimen in de teller en de noemer niet kent. I is een identifier, DP en DP' zijn (verschillende) domeinpseudoniemen, TP en TP' zijn (verschillende) transferpseudoniemen en PP is een projectpseudoniem.

Men kan aantonen dat aan de verschillende eigenschappen gegeven in sectie 3.2 voldaan is. Intuïtief is wel snel duidelijk dat onlinkbare pseudoniemen afgeleid uit dezelfde identifier slechts op het niveau van het project weer samenvallen: figuur 4 toont dus aan hoe twee verschillende domeinpseudoniemen (voor dezelfde identifier) op projectniveau herleid worden tot hetzelfde projectpseudoniem.

Dankzij de onderliggende assumptie van elliptische krommen is het onhaalbaar om uit een gekend pseudoniem de gebruikte sleutel te weten te komen. Daardoor is het eveneens onhaalbaar om de identifier of het pseudoniem waaruit het gekende pseudoniem afgeleid is te weten te komen.



Figuur 4. Realisatie m.b.v. elliptische krommen

Stel, bijvoorbeeld, dat uw I afgeleid is uit uw INSZ-nummer, dat uw jaarlijks inkomen beheerd wordt door de ene overheidsorganisatie en uw burgerlijke staat door de andere. Uw jaarlijks inkomen komt in het ene domein terecht onder pseudoniem DP, terwijl uw burgelijke staat terechtkomt in het andere domein onder pseudoniem DP'. Iemand die toegang heeft tot beide domeinen kan onmogelijk deze persoonsgegevens van u aan elkaar linken. Pas wanneer zowel uw jaarlijks inkomen als uw burgerlijke staat in een project ingeladen worden, kunnen de persoonsgegevens op het niveau van het project aan elkaar gelinkt worden.

Op basis van dit principe zijn drie algoritmes uitgewerkt die de eerste, de eerste twee of alle drie van de onderstaande vereisten vervullen:

- *Linkbaarheid mits samenwerking*. Het project kan attribuutwaarden van dezelfde persoon, afkomstig uit verschillende domeinen, aan elkaar linken, maar enkel mits medewerking van de betrokken domeinen.
- *Minimum knowledge voor projecten.* Het project leert niet meer dan dat wat strikt noodzakelijk is. Dit impliceert dat 1) het project niets te weten komt over (persoons)records die niet betrokken zijn in het project en 2) dat het project voor elk afzonderlijk record niet meer te weten komt dan de minimaal noodzakelijke gegevens. Indien een project vals speelt, wordt dit gedetecteerd.
- *Zero-knowledge voor domeinen.* Een domein leert geen nieuwe persoonsgegevens over de records wanneer het betrokken is in de uitvoering van een algoritme. Het leert zelfs niet welke records betrokken zijn in een project. Bij twee van de drie algoritmes ‘lekt’ er inderdaad informatie naar de domeinen.

De complexiteit van de algoritmes groeit naarmate aan meer vereisten voldaan moet worden.

Op basis van elliptische krommen kunnen we dus een efficiënte pseudoniemconversiefuncie creëren. Er werden op basis van de pseudoniemconversiefunctie drie verschillende algoritmes uitgewerkt om persoonsgegevens afkomstig van verschillende domeinen te kruisen in een project.

4. Uitbreidingen

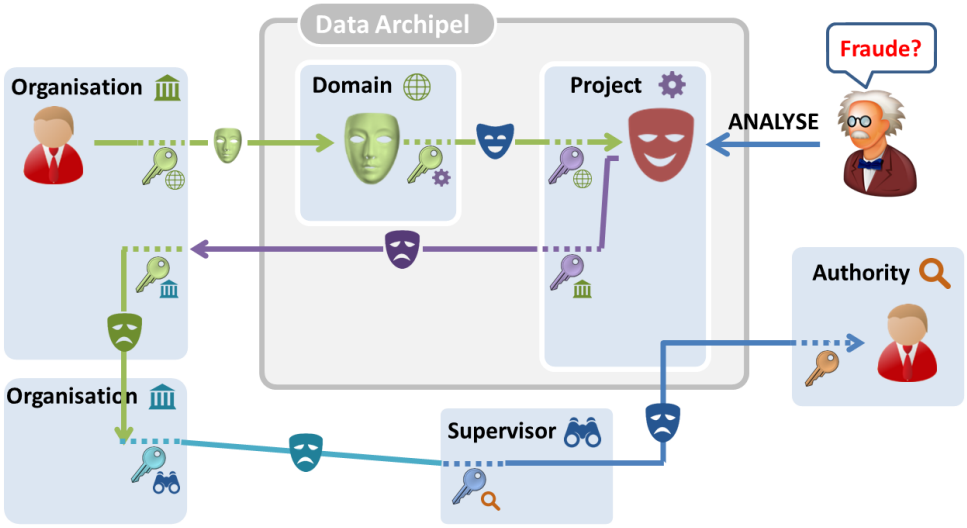
Verschillende uitbreidingen van de data archipel zijn mogelijk: gecontroleerde deanonimisatie, verschillende vormen van sleutelgeneratie en –distributie, transparantie naar de burger en virtualisatie. Deze sectie gaat hier dieper op in.

4.1. Gecontroleerde deanonimisatie

Onderzoekers analyseren de gekruiste data in projecten. In specifieke situaties is deanonimisatie (re-identificatie) van een pseudoniem vereist, bijvoorbeeld in de context van fraudebestrijding. Tegelijkertijd willen we dat niet de onderzoeker zelf, maar enkel een vertrouwde autoriteit, zoals een rechtbank of politie, in staat is deze deanonimisatie te doen. Soms is bijkomend de goedkeuring wenselijk van

één of meerdere andere partijen (vb. een toezichthouder zoals de privacycommissie).

De data archipel laat toe dat nul, één of meer partijen, naast het project en de vertrouwde autoriteit, hun medewerking moeten verlenen bij een deanonimisatieprocedure *per individu*. Elk van deze partijen moet daarbij een pseudoniemconversie doen met een specifieke sleutel voordat gedeanonimiseerd kan worden door de vertrouwde autoriteit. Daarbij wordt hetzelfde principe gebruikt als in sectie 3.3, maar dan in tegengestelde richting. Dit principe wordt geïllustreerd in figuur 5, waar drie entiteiten hun goedkeuring moeten geven vooraleer de vertrouwde autoriteit de identiteit van de betrokken burger te weten kan komen. De partijen die hun goedkeuring moeten geven vooraleer deanonimisatie mogelijk is kunnen aanleverende organisaties zijn of externe partijen (vb. een externe toezichthouder zoals de privacycommissie). Geen enkele van deze partijen leert informatie over de identiteit van de betrokken burger. Ze komen enkel te weten dat er vanuit een specifiek project een deanonimisatieaanvraag voor één burger is.



Figuur 5: Deanonimisatiepad (bestaande uit de droevige maskers) waarbij twee organisaties en een toezichthouder (vb. privacycommissie) hun medewerking moeten verlenen voordat een autoriteit een specifiek pseudoniem kan deanonimiseren.

De voorgestelde aanpak is erg flexibel en is compatibel met de wettelijke vereiste van proportionaliteit aangezien geen enkele partij zomaar om het even welk pseudoniem kan deanonimiseren.

4.2. Sleutelgeneratie en -distributie

Er zijn bepaalde relaties tussen de sleutels. Er is dus coördinatie vereist bij de generatie ervan voor de verschillende deelnemers en eilanden. De data archipel laat drie verschillende aanpakken toe:

- *Eén vertrouwde dealer.* De meest eenvoudige aanpak vereist één centrale dealer die alle sleutels genereert en deze vervolgens aan de juiste entiteit bezorgt (over een beveiligd kanaal). Het nadeel is uiteraard dat deze dealer alle geheimen kent en dat deze aanpak dus een grote mate van vertrouwen in de dealer vereist. Anderzijds kan dit een privépartij zijn, die losstaat van de overheid, waardoor de burgers een

grotere graad van vertrouwen kunnen krijgen in de data archipel ('de overheid' kent immers niet alle sleutels).

- *Meerdere vertrouwde dealers.* In plaats van één vertrouwde dealer kan gebruik gemaakt worden van meerdere vertrouwde dealers, die elk een deel van de sleutel genereren en veilig aan de juiste entiteit bezorgen. De ontvangende entiteit dient de verschillende sleuteldelen dan lokaal te combineren. Enkel wanneer alle vertrouwde dealers samenspannen, kunnen ze de sleutels te weten komen. Een vertrouwde dealer zou een overheidsinstantie kunnen zijn en een andere, bijvoorbeeld een private, speler. Het nadeel hier is de extra kost voor het opzetten en onderhouden van meerdere dealers.
- *Geen vertrouwde dealers.* In de laatste aanpak genereert elke organisatie en elk project zelf een uniek geheim. Vervolgens heeft het uitvoeren van specifieke protocollen tussen meerdere entiteiten als resultaat dat elke entiteit de nodige sleutels bezit, zonder dat er gebruik gemaakt wordt van een centrale autoriteit. Het vertrouwen is dus gedistribueerd, meer bepaald over de betrokken organisaties en het betrokken project. Het gevaar hier is opnieuw dat de burger de verschillende overheidsinstellingen als één geheel beschouwt en zo dus veronderstelt dat 'de overheid' opnieuw alle sleutels kent. De meerkost van de extra autoriteit(en) vervalt echter.

Tabel 1 vergelijkt de drie aanpakken. Naarmate er meer extra dealers zijn, stijgt de infrastructuurkost (1). Naarmate de sleutelgeneratie meer gecentraliseerd gebeurt is er een hogere graad van vertrouwen vereist in de entiteit of entiteiten die verantwoordelijk zijn voor deze sleutelgeneratie (2). Hoe meer partijen betrokken zijn bij de sleutelgeneratie, hoe groter de kans dat er ergens een sleutel lekt (3), maar daar staat wel tegenover dat het lekken van een sleutel minder verregaande consequenties heeft (4)

Vertrouwde dealers	**1**	**>1**	**0**
1. Kosten infrastructuur	Midden	Hoog	Laag
2. Vereist vertrouwen in dealer	Hoog	Midden	Laag
3. Kans op lek sleutel	Laag	Midden	Hoog
4. Impact bij lek sleutel	Hoog	Midden	Laag

Tabel 1. Vergelijking van de drie aanpakken voor sleutelgeneratie en -distributie

4.3. Transparantie naar de burger

Een aanleverende organisatie weet niet welke burgers betrokken zijn in een project en een project kent de identiteit niet van de betrokken burgers. Er is dus geen enkele overheidsorganisatie die een overzicht heeft van welke burgers er in welke project betrokken zijn.

Toch neemt dit niet weg dat aan de wettelijke vereiste van transparantie naar de burger toe kan voldaan worden. Dit kan door het publiceren van de beschrijvingen van alle (actieve en inactieve) projecten. De burger kan deze downloaden en op zijn lokale computer doorzoeken. Bij het geven van eigenschappen zoals geboortedatum, loon en postcode krijgt de burger de projecten te zien waarin gegevens van burgers die aan deze eigenschappen voldoen potentieel verwerkt

werden. Vervolgens kan de burger de gegeven zoekopdracht verder verfijnen door bijkomende eigenschappen te geven (drill-down).

Dus hoewel geen enkele overheidsinstelling weet welke burger er in welk project betrokken is, kan een burger dit toch over zichzelf en zijn naasten te weten komen door de projecten te filteren op attribuutwaarden zoals postcode.

4.4. Virtualisatie

In Figuur 1 is te zien dat gegevens gerepliceerd worden. Dezelfde persoonsgegevens kunnen zich bevinden op het niveau van de organisatie, het domein en het project. Hoewel dit het meest efficiënt toelaat om data te kruisen in een project, is het niet steeds de meest optimale aanpak. Replicatie impliceert immers een groter risico op inconsistenties en data breaches.

Daarom werd het concept van virtuele domeinen ingevoerd, waarbij de data zich voor projecten in een domein lijkt te bevinden, maar in werkelijkheid nog steeds effectief (en enkel) bij de organisatie staat. Wanneer een project data van verschillende domeinen kruist, worden de data door elke betrokken organisatie uit haar database geëxtraheerd en via het domein doorgestuurd naar het project. Dit resulteert echter in een extra belasting op de operationele systemen.

5. Proof-of-Concept

Een proof-of-concept werd ontwikkeld die de meest essentiële functionaliteit van de data archipel demonstreert. Bij voldoende vraag kan dit verder uitgebreid worden.

Figuur 6 in appendix A toont een screenshot waarbij de twee meest linkse vakken overheidsorganisaties zijn die persoonsgegevens beheren, gekoppeld aan een identifier. De twee middelste vakken zijn domeinen, waar persoonsgegevens onder domeinpseudoniemen bewaard worden en de rechtse vakken zijn twee projecten, waar data uit de twee domeinen gekruist wordt. In de organisaties is de burger gekend onder zijn INSZ-nummer, terwijl in elk van de domeinen en projecten, deze burger onder een ander pseudodniem gekend is.

> In appendix A is Marion Peeters door twee organisaties gekend met naam en INSZ-nummer. De ene organisatie kent voor elke burger zijn postcode en geboortedatum, de andere de professionele status en het exacte loon. Elk van de organisties stuurt deze persoonsgegevens door naar zijn domein. In beide domeinen is Marion Peeters nu enkel nog gekend onder twee verschillende domainpseudoniemen; de naam en het INSZ-nummer zijn niet langer aanwezig. In elk van de twee projecten wordt data afkomstig uit beide domeinen gekruist, wat enkel mogelijk is m.b.v. de cryptografische sleutels. Marion Peeters is in beide projecten gekend onder opnieuw andere pseudoniemen. Bemerk ook dat in beide projecten de attributen geboortedatum en domicilie verschillende gradaties van detail hebben.

De proof-of-concept toont niet enkel aan dat het theoretisch model ook in de praktijk werkt, maar het geeft ons ook al een ondergrens van de te verwachten performantie. De testen werden uitgevoerd op een laptop. Op een volwaardige server zullen de resultaten dus een pak beter zijn.

Tabel 2, rechts, toont de performantie van de pseudoniemconversiefunctie m.b.v. elliptische krommen. Gelijkaardige operaties m.b.v. RSA vergen een pak meer tijd (tabel 2, links). Eenzelfde rij komt overeen met eenzelfde niveau van veiligheid.

Een sleutellengte van 192 bits in elliptische krommen komt overeen met een sleutellengte van 1536 bits m.b.v. RSA. We zien trouwens dat de sleutellengtes voor elliptische krommen lineair stijgen, terwijl het voor RSA exponentieel is.

RSA		
Sleutel-lengte	*Duur 1 operatie*	*Operaties per uur*
1536 bits	58ms	62070
2048 bits	135ms	26700
3072 bits	440ms	8180

ECC (Elliptische krommen crypto)		
Sleutel-lengte	*Duur 1 operatie*	*Operaties per uur*
192	0,4ms	9 miljoen
224	0,6ms	9 miljoen
256	0,7-0,8ms	4-5 miljoen

Tabel 2. Performantie pseudoniemconversiefunctie op PC Windows 7 Enterprise (64bit) op één 2,66Ghz Intel i5 core

6. Uitdagingen & beperkingen

Een aantal technische uitdagingen blijft voorlopig bestaan. Uitdagingen 1, 2 en 3 passen in de **door wetenschappers verdedigde** [S15] **filosofie, dat de berekeningen naar de data gehaald moeten worden en niet andersom**. Het vierde puntje is een inherente beperking en het vijfde, ten slotte, is algemener en gaat veel ruimer dan de context van deze tekst. Deze tekst gaat niet in op de financiële consequenties, wat in de praktijk allicht de grootste uitdaging is.

1. Gedistribueerd uitvoeren van centraal geschreven scripts. Vandaag worden vaak persoonsgegevens uit verschillende bronnen op een centrale locatie samengebracht en gekruist. Vervolgens wordt een centraal script uitgevoerd dat op deze gegevens een filter toepast, wat resulteert in een set met beperktere gegevens die dan door de onderzoeker geanalyseerd kan worden. Zo kan de exacte geboortedatum vervangen worden door een leeftijdscategorie. Dergelijke filterscripts zullen nodig blijven om op een flexibele manier te kunnen antwoorden op de diverse vragen van onderzoekers waarbij enkel de minimaal noodzakelijke data in het project terechtkomen. Idealiter is het daarbij mogelijk dat de scripts nog steeds centraal geschreven worden, maar dat het wel uit meerdere modules bestaat, waarbij elk betrokken domein zijn module ontvangt en uitvoert en vervolgens het resultaat naar het project stuurt. Het filterscript wordt dus centraal geschreven maar decentraal uitgevoerd. Dit past in de filosofie om de berekeningen dichter bij de data te brengen. Die filosofie wordt momenteel door verschillende grote spelers op de markt gevolgd.

2. Attributen afgeleid uit meerdere domeinen. Soms kan een specifieke attribuutwaarde voor een project enkel gegenereerd worden uit persoonsgegevens die verspreid zijn over meerdere domeinen. Een project kan bijvoorbeeld enkel geïnteresseerd zijn in personen waarvan het loon, dat bewaard wordt door het ene domein, plus een premie, die gekend is door een ander domein, een bepaald bedrag overschrijdt ($domain_0.wage + domain_1.benefits > amount$). Of een project kan financiële gegevens nodig hebben over huishoudens, wat slechts afgeleid kan worden door het combineren van fiscale en familiale data over burgers, die door verschillende organisaties beheerd worden. *De uitdaging is dus om één attribuut af te leiden uit meerdere, verspreide attributen.* Is dit mogelijk zonder eerst veel gegevens ergens centraal te verzamelen en vervolgens er een filter op toe te passen?

3. Small-cell-probleem. Hoewel het small-cell-probleem op zich onoplosbaar is, kan de noodzaak om volledige records aan een externe partij prijs te geven gereduceerd worden. Dit kan door **het aanbieden van analytics tools in de data archipel zelf**. Het principe om analytics tools op het analytics platform zelf aan te bieden wordt door researchers verdedigd omwille van privacyredenen [S15] en

wordt reeds door enkele vendors aangeboden omwille van performantieredenen. Verder vereist een dergelijke aanpak degelijke security met o.a. toegangscontrole, monitoring en policies (wat mag de onderzoeker doen en wat mag niet). Door zo de berekeningen naar de data te halen i.p.v. andersom kan nagegaan en zelfs afgedwongen worden dat de onderzoeker niets doet met de data dat hij niet verondersteld is te doen, wat het principe van finaliteit ten goede komt. *Differential privacy* [T14] wordt gezien als één van de nieuwe tools om de berekeningen naar de data te halen. Voor testdoeleinden kan de onderzoeker eventueel toegang krijgen tot een kleine set records, die wel alle ruwe data bevat die door het project gekend is. Een meer drastische aanpak is de *safe room*, waarbij de onderzoeker slechts toegang heeft tot de gegevens in een project wanneer hij zich fysiek in een afgeschermde ruimte bevindt.

4. Domeinbeheerder ≠ projectbeheerder. Een belangrijke, inherente beperking is dat een projectbeheerder verschillend moet zijn van de beheerders van de aanleverende domeinen. Zoniet kunnen alle persoonsgegevens in een project gelinkt worden aan de persoonsgegevens in het domein. Beiden kennen namelijk de transferpseudoniemen.

5. Gestandardiseerde definities. Verschillende organisaties gebruiken afwijkende definities voor schijnbaar dezelfde attributen. De term werkloos kan bij organisatie 1 iets anders betekenen dan bij organisatie 2. Dit is een meer algemeen probleem dat niet in ons onderzoek behandeld wordt.

Samengevat zijn er nog een aantal uitdagingen die extra onderzoek vereisen. Toch lijken er geen blokkerende factoren te zijn.

7. Conclusie

De huidige Europse privacyrichtlijn en de toekomstige verordening plaatsen serieuze uitdagingen bij data-analyseprojecten op persoonsgegevens, in het bijzonder wanneer deze persoonsgegevens afkomstig zijn uit verschillende bronnen. Daarom werd de data archipel voorgesteld die de spanning tussen privacywetgeving enerzijds en analyse op gekruiste persoonsgegevens anderzijds met elkaar tracht te verzoenen, door o.a. de controle op de persoonsgegevens terug te geven aan de data controllers (de (overheids)bedrijven die verantwoordelijk zijn voor de persoonsgegevens). Deze beslissen zelf, na advies van hun juridische dienst, of ze al dan niet persoonsgegevens verschaffen aan een specifiek analyseproject. Ook ligt het veiligheidsniveau hoger dan bij een traditionele, gecentraliseerde aanpak, wat indirect de privacy van de burgers ten goede komt.

Vanuit een technisch standpunt is het combineren van persoonsgegevens voor een analyseproject vlot mogelijk. Aangezien de voorgestelde algoritmes enkel uitgevoerd worden wanneer gegevens gekruist worden, is er geen impact op de performantie tijdens de eigenlijke analyse van de gegevens, wanneer performantie cruciaal is.

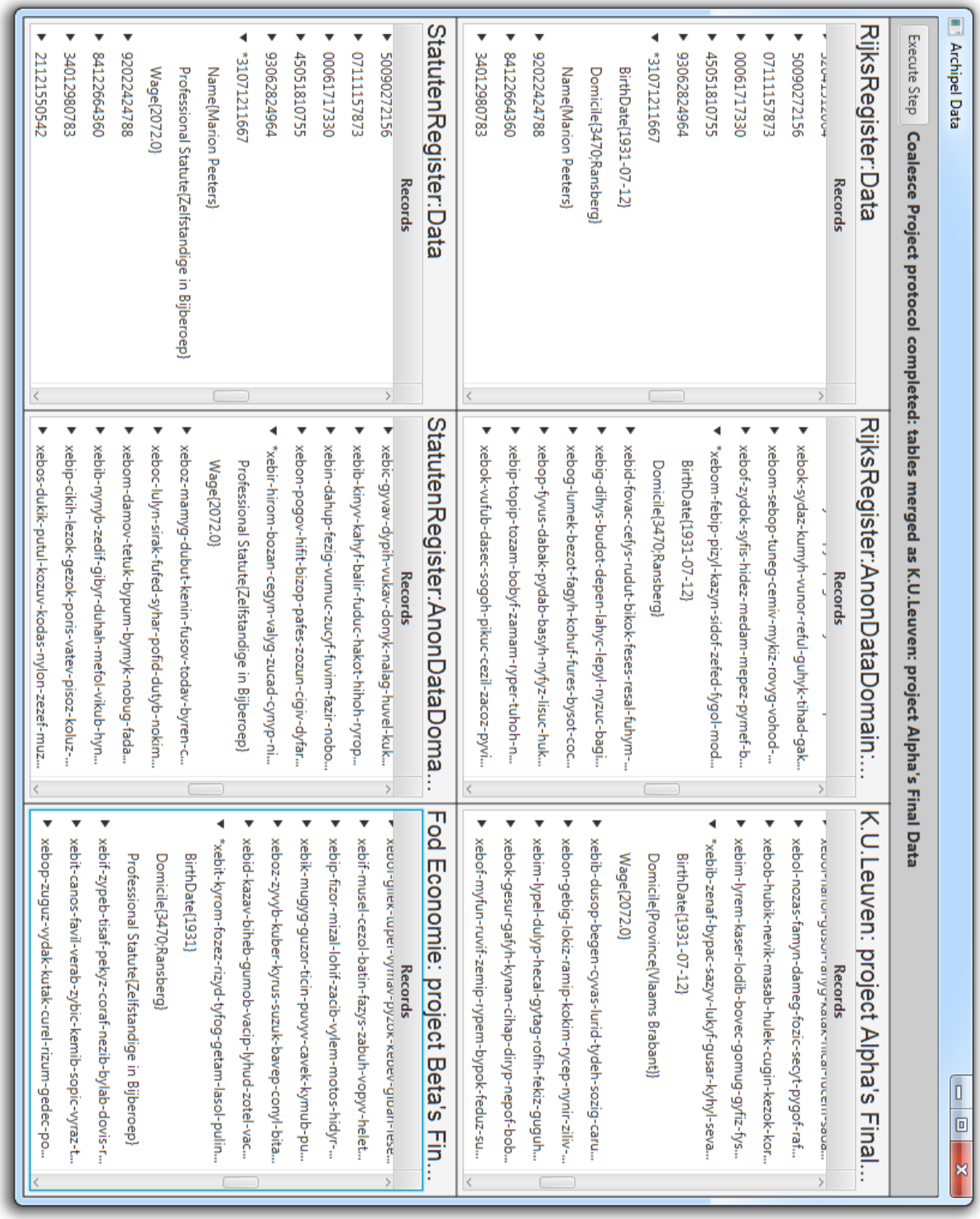
De voorgestelde aanpak is flexibel. Deanonimisatie door een vertrouwde autoriteit vereist de medewerking van nul, één of meerdere gekozen partijen, naast het project en de autoriteit. Sleutelgeneratie en -distributie kan gebeuren via één autoriteit, via meerdere autoriteiten of gedecentraliseerd. Verder zijn verschillende gradaties van virtualisatie mogelijk en wordt abstractie gemaakt van de onderliggende fysieke infrastructuur. Het sturen van updates vanuit de domeinen naar de projecten is mogelijk door de algoritmes opnieuw uit te voeren.

Samengevat levert de data archipel op verschillende vlakken een meerwaarde. Het helpt om analyseprojecten met persoonsgegevens afkomstig uit verschillende

bronnen te laten voldoen aan de wettelijke vereisten van transparantie, proportionaliteit, finaliteit en information security practices. Samen met de geleverde flexibiliteit heeft de data archipel een sterke toegevoegde waarde t.o.v. de huidige manier van werken, al is nog extra onderzoek vereist om de nog bestaande uitdagingen aan te pakken.

Hoewel het niet evident is om de data archipel volledig te realiseren, zijn er een aantal ideeën die ook buiten de data archipel toegepast kunnen worden in een data analysecontext. Het concept van de data archipel laat ons in elk geval zien in welke richting we kunnen evolueren met betrekking tot het kruisen van persoonsgegevens voor analysedoeleinden.

Appendix A



Figuur 6: Screenshot Proof-of-Concept. De twee linkse vakken zijn overheidsorganisaties die persoonsgegevens beheren, gekoppeld aan een identifier. De twee middelste vakken zijn domeinen, waar persoonsgegevens onder domeinpseudoniemen bewaard worden en de rechtse vakken zijn twee projecten, waar data uit de twee domeinen gekruist wordt en gekend is onder projectpseudoniemen.

Referenties

[AV07] Arvind Narayanan, Vitaly Shmatikov. *How To Break Anonymity of the Netflix Prize Dataset*. Cornell University Library. 22 november 2007.
http://arxiv.org/abs/cs/0610105

[B92] *Wet van 8 december 1992 tot bescherming van de persoonlijke levensfeer ten opzichte van de verwerking van persoonsgegevens*. 1999.
http://www.e-privacy.be/privacywet.pdf

[BT06] Michael Barbaro and Tom Zeller, Jr., *A Face is Exposed for AOL Searcher No. 4417749*, The New York Times, 9 augustus 2006.
http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all

[M13] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel. *Unique in the Crowd: The privacy bounds of human mobility*. 25 maart 2013. Scientific Reports 3, Article number: 1376.
http://www.nature.com/srep/2013/130325/srep01376/fig_tab/srep01376_F1.html

[M15] Mark Hosenball. *U.S. has yet to notify 21.5 million data breach victims: officials*. Juli 2015.
http://www.reuters.com/article/2015/07/14/us-cybersecurity-usa-notification-idUSKCN0PO2SE20150714

[M15a] Tania Martin, Smals research. *Elliptic Curve Cryptography for dummies 1: introduction*. 25 februari 2015.
https://www.smalsresearch.be/elliptic-curve-cryptography-tutoriel1/

[M15b] Tania Martin, Smals Research. *Elliptic Curve Cryptography for dummies 2: en pratique pour la cryptographie*. 12 Augustus 2015
https://www.smalsresearch.be/elliptic-curve-cryptography-tutoriel2/

[NS09] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset). 2006
arxiv.org/pdf/cs/0610105

[O09] Paul Ohm. *Broken Promises of Privacy: Responding to the surprising failure of anonymisation*. UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. 13 augustus 2009
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006

[PCAST14] President's Council of Advisors on Science and Technology (PCAST). *Big Data and Privacy: A Technological Perspective.* Mei 2014.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

[S15] John Bohannon, Credit card study blows holes in anonymity. Science Magazine, 20 january 2015. Vol 347 issue 6221. P 468
http://www.sciencemagazinedigital.org/sciencemagazine/30_january_2015?folio=468

[T14] Anthony Tockar. Differential Privacy: The Basics. Neustar Research. 8 september 2014.
http://research.neustar.biz/2014/09/08/differential-privacy-the-basics/

[V15] Kristof Verslype, Smals Research. *Big data & krakend ijs onder anonimisatie*. 12 mei 2015.
https://www.smalsresearch.be/big-data-krakend-ijs-onder-anonimisatie/

[VD16] Kristof Verslype & Bart De Decker. *Data Archipelago – Reconciling privacy with analytics on multi-source PII.* 2016
SUBMITTED FOR: Proceedings on Privacy Enhancing Technologies 2016