METHODE VOOR HET VERMIJDEN VAN VERMELDINGEN VAN EXACTE DATUMS IN DATASETS DIE AAN ONDERZOEKERS TER BESCHIKKING WORDEN GESTELD

Context

In de verwerking van aanvragen voor het ter beschikking stellen van datasets aan onderzoekers wordt het Informatieveiligheidscomité regelmatig geconfronteerd met de vraag om exacte datums te verkrijgen. Deze datums dienen beschouwd te worden als een “quasi-identifier” waardoor de kans op identificatie van een betrokkene reëel wordt, zelfs wanneer de identifiers gepseudonimiseerd zijn of niet langer aanwezig zijn in de dataset.

Wanneer mogelijk, zal de aanvraag voor informatie aan het Informatieveiligheidscomité aangepast worden zodat niet langer datums maar wel perioden opgevraagd worden. Toch zal dit niet steeds een werkbare oplossing zijn voor de onderzoeker omdat deze vaak de datums zal gebruiken om duurtijden te berekenen.

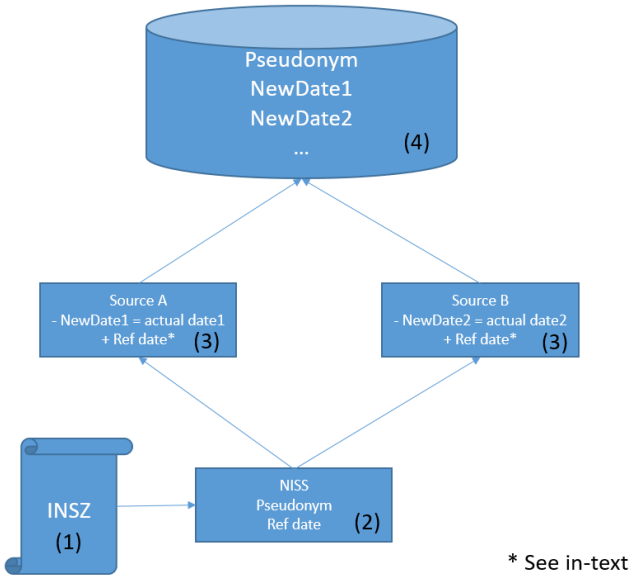
De complexiteit van het antwoord verhoogt nog wanneer de datums afkomstig zijn vanuit 2 verschillende authentieke bronnen. Op dat moment zou een TTP de informatie moeten interpreteren en berekeningen uitvoeren wat niet noodzakelijk tot de opdracht van de TTP behoort en de vrijheid van de onderzoeker sterk zou beperken.

Om hierop een passend antwoord te kunnen geven is een voorstel uitgewerkt waarbij niet langer exacte datums zouden moeten doorgegeven worden door een authentieke bron, maar waar de datums herberekend worden t.o.v. een vaste referentiedatum die willekeurig wordt bepaald per betrokkene van wie gegevens in de dataset zullen voorkomen. Doordat de referentiedatum niet in de finale dataset voorkomt, zal het niet langer mogelijk zijn de exacte datums terug te rekenen op basis van deze dataset.

Op deze manier kan op eenvoudige en kost-efficiënte wijze een oplossing geboden worden voor de meeste aanvragen waarbij exacte datums noodzakelijk geacht worden.

Voorstel en technische uitwerking

Het voorstel steunt op het volgende principe:



0. Het tijdsdomein voor het onderzoek wordt bepaald en het aantal dagen (MaxDays) wordt berekend.
1. Een lijst van INSZ wordt opgesteld om de steekproef van het onderzoek te bepalen.
2. Een TTP zal voor elk van deze INSZ:
    a. een pseudoniem creëren en
    b. een willekeurig gekozen natuurlijk getal[1] (Ref date) bepalen kleiner dan het aantal dagen van het tijdsdomein.
3. Elke authentieke bron zal zijn NewDates herberekenen aan de hand van volgende formule[2]:
    a. NewDate = (Startdatum Tijdsdomein) + ((Reële datum-Startdatum tijdsdomein) + (Ref date)) mod (MaxDays)
4. Voor het berekenen van duurtijden dient de onderzoeker volgende formule toe te passen:
    a. Duration = (NewDate2 – NewDate1 + MaxDays) mod (MaxDays)

Bepalen van het tijdsdomein

Als eerste stap in deze oplossing wordt een tijdsdomein bepaald dat gebruikt zal worden voor de willekeurige verschuiving van de datums per betrokkene.

Algemeen kan gesteld worden dat het tijdsdomein groter of gelijk moet zijn dan de periode waarop het onderzoek betrekking heeft, vermeerderd met de hoogste duurtijd tussen de startperiode en elk event.[3]

Dit tijdsdomein dient te worden bepaald door de aanvrager en kan berekend worden op basis van input die bekomen kan worden bij authentieke bronnen. De authentieke bronnen zullen op dat moment een voldoende marge moeten voorzien zodat deze input zelf geen exacte datum zal zijn.

---

[1] Bv. SecureRandom functie van een computer waarbij zoveel als mogelijk een uniforme verdeling wordt bekomen.
[2] mod : modulo berekening geeft het restgetal na deling van een geheel getal door een ander geheel getal
[3] Bv. een onderzoek waarbij de incidentiedatum als selectiecriterium wordt genomen en de periode waarop het onderzoek betrekking heeft is 3 jaar. Wanneer er datums van evenementen in de dataset zijn die voor een betrokken tot 1 jaar na incidentiedatum kunnen uitlopen, dan zou het tijdsdomein (3 + 1) jaar zijn en dus 1461 dagen.

In het geval dat er geen input van de authentieke bron kan bekomen worden om het tijdsdomein te berekenen, kan het tijdsdomein lopen van de start van de periode waarop het onderzoek betrekking heeft tot de dag van de extractie van gegevens door de authentieke bronnen.

Risico's

Volgende risico's werden overwogen en de beschermingsmethoden beschreven:

1. Bepalen van exacte datums aan de hand van de referentiedatums:
    a. De finale dataset mag geen informatie bevatten over de referentiedatum zodat deze berekening niet kan gebeuren. De ontvangende partij mag ook op geen andere manier over deze referentiedatums beschikken.
    b. Dit voorstel voorziet in het vermijden van het doorgeven van exacte datums voor het berekenen van duurtijden. Omdat dit relatieve berekeningen zijn is de referentiedatum niet nodig voor een correcte uitkomst.
2. Schatten van reële datums aan de hand van de NewDates:
    a. Door het gebruik van de modulo berekening wordt elke reële datum geprojecteerd op een willekeurige plaats in het tijdsdomein dat bepaald werd bij de aanvraag. De positie van de Newdate is de som van de referentiedatum en de reële datum. Doordat de moduloberekening wordt toegepast, valt het niet te achterhalen of de som van deze 2 data kleiner of groter is dan het maximum aantal dagen in het bepaalde domein.
3. Verlies van kwaliteit van informatie door verschuiving van datums:
    a. De berekeningen die voorzien zijn op NewDates zijn duurtijden. Doordat deze bewerking relatief is, en de verschuiving voor beide datums dezelfde, is er geen kwaliteitsverlies in deze berekening. De onderzoeksinstelling zal zijn algoritmes wel dienen aan te passen aan dit voorstel.
4. Verlies van integriteit van de bekomen informatie door het toepassen van de modulo-berekening:
    a. Wanneer een duurtijd tussen bv. incidentie en een geassocieerd evenement absoluut groter zou zijn dan het tijdsdomein, dan riskeert de berekende duurtijd een veelvoud van de duurtijd van het tijdsdomein te verschillen. Dit moet goed nagegaan worden bij de bepaling van het tijdsdomein.

Voorbeeld

Ter illustratie wordt een voorbeeld uitgewerkt met volgende willekeurig bepaalde parameters.

Startdatum onderzoeksperiode :	01/01/2010 (00:00)
Einddatum onderzoeksperiode :	31/12/2019 (23:59)
Hoogste duurtijd tussen een tussen eerste en laatste evenement voor elke betrokkene :	2 jaar
Reële datum1	15/2/2016
Reële datum2	13/07/2018
Reële datum3	20/10/2020

Berekening tijdsdomein

Het tijdsdomein is de som van de duurtijd van de onderzoeksperiode en de hoogste duurtijd tussen een tussen eerste en laatste evenement voor elke betrokkene.

In dit geval is dit:

Duurtijd onderzoeksperiode :	3652 dagen (10 jaar)
Hoogste duurtijd tussen een tussen eerste en laatste evenement voor elke betrokkene :	732 dagen[4] (2 jaar)
Duurtijd tijdsdomein (MaxDays) :	4384 dagen

Voor het tijdsdomein geldt dan:

Start :	01/01/2010 (00:00)
Einde :	2/01/2022 (00:00)

Berekening duurtijd

De duurtijden, berekend o.b.v. de echte datums zijn:

Duurtijd Reële Datum1 – Reële Datum2 :	879 dagen
Duurtijd Reële Datum1 – Reële Datum 3 :	1709 dagen

Voor het bepalen van de NewDate's wordt eerst een referentiedatum gekozen. Dit gebeurt aan de hand van een willekeurig gekozen natuurlijk getal kleiner dan MaxDays. Voor dit voorbeeld wordt RefDate = **956** (14/8/2012) gekozen.

NewDate = (Startdatum Tijdsdomein) + ((Reële datum tijdsdomein) + (Ref date)) mod (MaxDays). Dit geeft volgende resultaten:

NewDate1	(01/01/2010) + (((15/02/2016) – (01/01/2010)) + 956)mod(4384)) dagen= (01/01/2010) + ((2236 + 956)mod(4384)) = (01/01/2010) + 3192 = **28/9/2018**
NewDate2	(01/01/2010) + (((13/07/2018) – (01/01/2010)) + 956)mod(4384)) dagen= (01/01/2010) + ((3115 + 956)mod(4384)) = (01/01/2010) + 4071 = **23/02/2021**
NewDate3	(01/01/2010) + (((20/10/2020) – (01/01/2010)) + 956)mod(4384)) dagen= (01/01/2010) + ((3945 + 956)mod(4384)) = (01/01/2010) + ((4901)mod(4384)) =

[4] Om de berekening eenvoudig te houden werd hier gekozen voor 366 dagen per jaar. Op deze manier is men zeker dat het effect van een schrikkeljaar niet gemist wordt.

(01/01/2010) + 517 = **02/06/2011**

Voor de berekening van de effectieve duurtijden o.b.v. de NewDates wordt volgende formule gebruikt:

Duration = (NewDate2 – NewDate1 + MaxDays) mod (MaxDays)

Voor dit voorbeeld geeft dit :

Duurtijd NewDate1 – NewDate2 : ((23/02/2021 – 28/09/2018)+ 4384)mod(4384) =
(879 + 4384)mod(4384) =
(5363)mod (4384) =
879 dagen

Duurtijd NewDate1 – NewDate3 : ((02/06/2011 – 28/09/2018)+ 4384)mod(4384) =
(-2675 + 4384)mod(4384) =
(1709)mod (4384) =
1709 dagen

Grafisch voorgesteld geeft dit het volgende beeld: