Linked open data & AVG: nik

In dit artikel verkennen we de opkomst van linked open data, en daarbij het combineren van data, in het licht van privacy. We kijken dan met name naar de Algemene verordening gegevensbescherming (AVG).

Door Erwin Folmer en Mathieu Paapst

Interoperabiliteit, de mate waarin systemen/organisaties met elkaar kunnen samenwerken, is al decennia een groot streven. Standaarden op technisch, semantisch en organisatorisch niveau spelen daarin een grote rol. In de laatste jaren hebben de W3C-standaarden rond linked data een vlucht genomen, waardoor interoperabiliteit op technisch en semantisch niveau een steeds hoger niveau haalt. In essentie worden data meer en meer op het web gepubliceerd, zijn data eenvoudig koppelbaar en integraal bevraagbaar met andere data op het web. Een krachtig fenomeen, waar veel waarde mee te generen is, maar dan rijst ook de vraag of er geen misbruik van te maken is. Of specifieker: als al die data koppelbaar zijn en zo eenvoudig nieuwe inzichten worden vergaard, is privacy dan nog wel te garanderen?

In deze bijdrage willen we kort stilstaan bij de vraag of en hoe linked data zich verhouden tot privacy en of de AVG een afdoende bescherming kan vormen voor eventuele privacygevaren bij het toepassen van linked data.

Om een goede analyse te kunnen uitvoeren, is zuiverheid qua begrippen essentieel. Als we praten over 'linked data', dan zijn dat de linked data-standaarden toegepast in gesloten datacontext (of geen expliciet onderscheid). Als we 'linked open data' gebruiken, dan hebben we het specifiek over de linked data-standaarden toegepast op datasets met een open licentie. Met linked data worden data in een netwerk verbonden (in tegenstelling tot het publiceren van datasilo's), ook wel bekend onder de noemers 'web of data', 'semantisch web' en 'knowledge graph' (ook opgenomen in de Gartner Technology Hype Cycle 2018). Met een knowledge graph worden data integraal bevraagbaar en kunnen machines over data gaan redeneren om nieuwe inzichten boven tafel te krijgen. Niet voor niks maken de grote informatiebedrijven (Google, Facebook, enzovoort) allemaal gebruik van een eigen knowledge graph. In de context van LOD is er sprake van een open knowledge graph op het web. Een belangrijke 'verbinder' van data is locatie: nagenoeg hebben alle data wel iets van een geo/locatie-component in zich.

Als we spreken over privacy, dan dienen we dat te onderscheiden in: de relationele

privacy (de bescherming van je gezinsleven), de ruimtelijke privacy (de bescherming van wat je doet in je woning), en de informationele privacy (de bescherming van gegevens die direct of indirect iets zeggen over een persoon). Het is dat laatste waar de Algemene verordening gegevensbescherming (AVG) op van toepassing is.

Identificatie

De meest belangrijke definitie in de AVG is die van het begrip 'persoonsgegevens'. Het gaat daarbij om alle informatie over een geïdentificeerde of een direct danwel indirect identificeerbare persoon. Vooral de term 'indirect identificeerbaar' maakt dat allerlei data die op het eerste gezicht geen persoonsgegevens zijn, toch al heel snel onder de reikwijdte van de AVG kunnen vallen. Daarbij is bijvoorbeeld niet vereist dat alle informatie om een persoon te kunnen identificeren bij een en dezelfde persoon of organisatie berust. Zodra je kunt beschikken over al dan niet wettelijke middelen waarmee je data kunt koppelen aan gegevens afkomstig van een derde, en

s aan de hand, of toch wel?

zodoende identificatie tot stand zou kunnen brengen, zijn de data te beschouwen als persoonsgegevens. Terecht kan dan de vraag gesteld worden of het beschikbaar stellen van open datasets en een middel waarmee die datasets gekoppeld kunnen worden, geraakt gaat worden door de AVG.

Algemene analyse

Alvorens toe te komen aan het linken van data en de mogelijke gevolgen daarvan, dienen we eerst te kijken naar de AVG consequenties bij het beschikbaar stellen van datasets. Om onder de AVG te vallen, dient er sprake te zijn van mogelijke persoonsgegevens die in een bestand zijn opgenomen. Daarbij is het niet noodzakelijk dat er sprake is van geautomatiseerde verwerking. Bij het openstellen van een dataset is het al goed gebruik om te toetsen of het belang van de openbaarheid zwaarder weegt dan het belang van de bescherming van de persoonlijke levenssfeer. Het uitgangspunt is hier dat, wanneer er gegevens in de dataset staan over een geïdentificeerde persoon, zo'n dataset - wettelijke uitzonderingen daargelaten - niet openbaar wordt gemaakt door de houder van de dataset. Datzelfde is van toepassing indien er gegevens instaan die direct of indirect identificatie van de betrokkene door de houder tot stand kunnen brengen. Bijvoorbeeld door aanvullende gegevens bij derden op te vragen. Omdat de houder hierbij dus van tevoren moet afwegen of zij over de hiertoe benodigde (wettelijke) middelen kan beschikken, is dit een lastige afweging.

Nu kan het natuurlijk zo zijn dat er een dataset na bovenstaande toetsing openbaar wordt gemaakt en een derde partij deze dataset weet te koppelen met andere open datasets. In de ideale wereld zijn ook die andere datasets van tevoren door de desbetreffende houders getoetst op bovengenoemde uitgangspunten en verandert het aan elkaar linken van die verschillende open datasets - die immers ieder afzonderlijk geen persoonsgegevens bevatten - niets. De AVG was en is daarop niet van toepassing en je komt daarom ook niet toe aan vraagstukken rondom een voor verwerking van persoonsgegevens benodigde grondslag, dataminimalisatie of doelbinding.

Die vlieger gaat echter niet op indien die derde partij zelf ook houder is van een dataset waarin wel persoonsgegevens zijn opgenomen en zij die bewuste dataset intern weet te koppelen aan gegevens uit een open dataset. Dat is een situatie die weliswaar in de dagelijkse praktijk zal kunnen voorkomen, maar die desondanks ten tijde van de beschikbaarstelling voor de houder van de open dataset niet te voorzien was. Het is daarom ook niet die houder die voor een dergelijke verwerking verantwoordelijk kan worden gesteld, maar juist de derde partij die voor de interne koppeling en het gebruik van de aldus verkregen extra informatie een grondslag dient te hebben zoals genoemd in de AVG.

Specifieke situaties

Voorgaande laat zien dat de basis in principe goed is. Ook bij vaak gebruikte hypothetische voorbeelden als een pedofiel die met behulp van meerdere open datasets kan uitzoeken welk zwembad in welke wijk ligt met de meeste jonge kinderen, of een inbreker die op basis van meerdere datasets een wijk kan bepalen met de hoogste woningwaarde en de hoogste gemiddelde leeftijd van de bewoners is de basis goed. De discussie of het vrij beschikbaar zijn van deze datasets heeft bijgedragen aan het plegen van het misdrijf, staat dus volledig los van de vraag of het beschikbaar stellen en het gebruik van de dataset in strijd is met de AVG. Ook als persoonsgegevens bijvoorbeeld in het online telefoonboek zijn opgenomen (op basis van toestemming van de betreffende persoon), dan is de gebruiker van die informatie (en niet de aanbieder van het online telefoonboek) fout bezig als hij deze persoonsinformatie koppelt aan andere data als daar geen grondslag voor is.

Hoe zit het dan met eventuele tussenpersonen, zoals partijen die het doorzoeken van verschillende datasets faciliteren? De AVG is daar helder over: op voorhand ben je als dataprovider (tussenpersoon) niet direct verantwoordelijk (6:196c BW) om alle data en toepassingen te controleren. Jurisprudentie rond bekende

zaken als Kazaa en Pirate Bay laten echter wel een in de tijd opschuivend beeld zien: van totaal geen aansprakelijkheid van tussenpersonen tot vormen van aansprakelijkheid indien de tussenpersoon op de hoogte is of kan zijn van het onrechtmatige karakter. Hier wordt het beeld al iets grijzer, maar er is meer. Neem de zwembaden: data hierover zijn zeer nuttig voor goede toepassingen en mogelijk ook zeer beperkt voor negatieve toepassingen. Blijkbaar is de afweging gemaakt en besloten tot open publicatie. Maar wat als die verhouding in impact nou verschuift? In de nieuwe situatie betreft het een data-object waarmee grotendeels alleen maar niet toegestane toepassingen bedacht kan worden. Dan lijkt het minder gewenst om deze data te publiceren, maar nog niet per definitie strafbaar.

Complexer wordt het als de 'foute toepasser' ergens een systeem gebruikt om de analyse te kunnen maken, waarbij de data vanuit een ander systeem/locatie gebruikt wordt. Een functionele scheiding tussen analyse en data. In termen van linked data: een zogenaamde federated query, die door de gebruiker uitgevoerd wordt op een SPARQL endpoint (data platform), dat vervolgens andere systemen gaat bevragen. Als eigenaar van het dataplatform weet (bijvoorbeeld op basis van logging) dat jouw platform wordt gebruikt voor de foute analyse, ook al ben je niet de dataleverancier, dan heb je de verplichting om deze toepassing eruit te filteren/blokkeren. Filtering (query blokkade) lijkt dan ook een must-have functionaliteit voor een linked data-platform.

Met linked data wordt dit wel erg makkelijk. Dan rijst de vraag: Hebben we het niet te makkelijk gemaakt? Dus staat linked open data niet haaks op privacy by design? Je zou privacy by design kunnen uitleggen als het ontwerpen en publiceren van datasets waarbij linking onmogelijk wordt gemaakt, zodat privacy by design is gegarandeerd. Dat staat wel haaks op de principes van linked data, waarbij datasets ontworpen worden op maximale herbruikbaarheid en verbondenheid.

Het is duidelijk dat persoonsgegevens niet gepubliceerd mogen worden, maar we verwachten ook een verschuiving naar een verbod op publicatie van andere open data, omdat de impact niet onverdeeld positief is (niet gelimiteerd tot alleen privacy risico's). In welke mate dit gaat schuiven, hangt af van de maatschappelijke normen en waarden: wat vinden we wel/niet acceptabel?

Een losstaand issue. Is het foutief verbinden van data, foutieve analyses (waardoor verkeerde conclusies getrokken kunnen worden) vergelijkbaar met fake news? Dit kan vervelend zijn en de concepten van linked data trachten dit te voorkomen door semantiek bij de data te leveren, maar uiteindelijk heeft iedereen recht op zijn eigen gebrek. linked open data levert echter wel transparantie in zowel de analyse als de data en dat kan wel helpen bij het eerder ontdekken van foute analyses.

Conclusies

In de basis levert de combinatie van AVG en linked data geen issues op; de 'grondslag' voor de toepasser is het fundament. Maar het ontslaat de data(platform)aanbieder niet om verantwoordelijk met data om te gaan. Er zit een spanningsveld bij linked open data en privacy by design. Dit spanningsveld zal zich ontwikkelen en verschuiven op basis van het maatschappelijke debat rond privacy en de spelregels van het publiceren van (linked) open data zullen veranderen onder druk van het maatschappelijke debat. Dit is een debat hoe open/transparant de samenleving moet zijn.

Er is ook een positieve kant: doordat er toch ook een verantwoordelijkheid ligt bij de data-aanbieder kan dit ook leiden dat data minder 'over de schutting' worden gepubliceerd (publiceren omdat het moet), maar dat de focus verschuift naar een 'begeleid gebruik' van data door de data-aanbieder. Daardoor ontstaan betere toepassingen en beter inzicht in het gebruik van de data, wat kan leiden tot beter datapublicatie, waardoor er meer waarde ontstaat. Oftewel: dan komt de beoogde levenscyclus van open data pas echt goed op gang.

Erwin Folmer is werkzaam bij Kadaster & Universiteit Twente. Erwin is bereikbaar via erwin.folmer@kadaster.nl.

Mathieu Paapst is werkzaam bij ICTRecht & Vrije Universiteit. Mathieu is bereikbaar via m.paapst@ictrecht.nl.

GIN-enquête

Vorige maand is er een enquête naar de leden van GIN verstuurd. Deze enquête bevat vragen over Geo-Info, hét magazine van GIN.

Wil je ook laten weten wat je van het magazine vindt? We bieden je de mogelijkheid de vragenlijst ook in te vullen. Via deze link kom je bij de vragenlijst: bit.ly/2JUHasW.

Namens de redactie van Geo-Info alvast hartelijk dank voor het invullen van de vragenlijst!