Research and Data Centre

Cahier 2024-21

Data lineage for the justice system

Scope, potentials, and directions

Samenvatting

Dutch institute for Justice and Security

Cahier 2024-21

Data lineage for the justice system

Scope, potentials, and directions

Samenvatting

Mortaza S. Bargh

Cahier
This series comprises overviews of studies carried out by or for the WODC Research and Data Centre. Inclusion in the series does not mean that the sheet's contents reflect the viewpoint of the Minister of Justice and Security.

Samenvatting

Data lineage voor het rechtsstelsel

Toepassingsgebied, potentieel en aanbevelingen

Probleemstelling

Probleem (context)

Data worden momenteel in een steeds sneller tempo gegenereerd, verzameld, gedeeld, geanalyseerd en verspreid. Als gevolg van deze ontwikkeling is er een toenemende belangstelling voor (en vraag naar) methoden om de beschikbare data bij elkaar te brengen, met behulp van (geavanceerde) algoritmen te analyseren en datagestuurde systemen te ontwikkelen om ons dagelijks leven te verlichten, toegevoegde waarde te creëren voor bedrijven, inzicht te geven in maatschappelijke fenomenen, en beleidsvormingsprocessen te sturen. De wijze waarop data worden verzameld, wat vaak gepaard gaat met subjectieve, partijdige, foutieve, gevoelige en stigmatiserende informatie over individuen, groepen en organisaties, en de wijze waarop algoritmen en datagestuurde systemen worden ontworpen, geïmplementeerd, geïnterpreteerd en (verkeerd) gebruikt, hebben grote invloed (of gaan dat hebben) op ons als individu, als groep en als maatschappij. Als een organisatie profijt wil hebben van data, moet zij dus alert zijn op de risico's van de data die worden gebruikt om persoonlijke, sociale of organisatorische voordelen te bieden. Vertrouwen scheppen in data is dan ook een absolute voorwaarde voor het respecteren van fundamentele mensenrechten zoals privacy, vrijheid, autonomie en waardigheid.

Ook op justitieel gebied, met name in het Nederlandse rechtsstelsel, worden in toenemende mate digitale technologieën en datagestuurde systemen toegepast. De informatiesystemen in het justitieel apparaat, die gegevens verzamelen, opslaan, delen en verwerken, zijn vaak fysiek verspreid, hebben veel losjes gekoppelde subsystemen en worden beheerd door verschillende organisaties (d.w.z. verspreid over veel administratieve domeinen). Om in deze setting data te gebruiken, moeten verschillende informatiebronnen met elkaar worden verbonden en moeten de data op een betrouwbare en verantwoorde manier worden geïntegreerd. Degenen die data delen (zoals gerechtelijke dienstverleners) moeten op de datagebruikers vertrouwen dat zij de data op verantwoorde wijze gebruiken, en degenen die data consumeren (zoals beleidsmakers) moeten er vertrouwen in hebben dat gegevensbronnen op verantwoorde wijze gegevens verzamelen en delen.

Vertrouwen in dataproductie en -gebruik vereist onder meer het terugdringen van de beveiligingsproblemen van grotendeels verspreide gegevensafhankelijke systemen, het managen van datakwaliteitsproblemen van losjes gekoppelde databronnen en het aanpakken van onrechtmatig en/of kwaadwillig gebruik van data en algoritmeresultaten. Gezien het belang van gegevensuitwisseling enerzijds en de toegenomen complexiteit van gegevensuitwisseling tussen (de vele) belanghebbenden en informatiesystemen anderzijds, is het noodzakelijk een passend data-ecosysteem tot stand te brengen. Een dergelijk data-ecosysteem vereist solide en effectieve data-

governance en effectief databeheer om de kwaliteit van data te waarborgen, de opslag en uitwisseling van data te beveiligen, de wisselwerking tussen concurrerende waarden (zoals datagebruik en dataprivacy) te optimaliseren, en de beginselen van vindbaarheid, toegankelijkheid, interoperabiliteit en herbruikbaarheid voor de data te operationaliseren.

Een efficiënte en effectieve data-governance/-management vereist onder andere kennis over de geschiedenis van data (hierna *data lineage*) en de herkomst van data (hierna *data provenance*). Data lineage verwijst naar het proces van traceren van de datastroom in de tijd, d.w.z. tijdens de levenscyclus/het traject van de data. Met metadata wordt een duidelijke beschrijving gegeven van de herkomst van de data, de datatransformaties langs het datatraject en de bestemming(en) van de data. Een soortgelijk geval van data lineage is data provenance, d.w.z. het bijhouden van de herkomst (de oorsprong) van data en de historische veranderingen ervan.

Data lineage wordt gezien als een essentieel instrument om de betrouwbaarheid van data te verbeteren. Het maakt bijvoorbeeld efficiënt beheer van datakwaliteit, datawijzigingen en datalevenscycli mogelijk. Een bekende metafoor die gebruikt wordt om de rol van data lineage bij het scheppen van vertrouwen in data te illustreren is het scheppen van vertrouwen in de voedzaamheid en gezondheid van een appel door kennis over hoe een appel wordt gekweekt, geoogst, vervoerd, opgeslagen, gedistribueerd en verkocht. Om vertrouwen te scheppen, zou men de relevante informatie (informatie over de ontstaansgeschiedenis) in elke fase van de levenscyclus van de appel (d.w.z. de route die de appel aflegt) in de toeleveringsketen kunnen bijhouden. In deze metafoor is het object in kwestie een fysiek object (product). Eenzelfde vorm van vertrouwen is ook belangrijk voor een digitaal object (bijv. een dataset, een digitale afbeelding, of een digitaal document). Zo worden burgers geconfronteerd met een steeds grotere hoeveelheid multimedia-content (oftewel data) in verschillende formaten, zoals afbeeldingen, video's, audio-opnamen en documenten. Deze content wordt vaak vermengd met desinformatie (bijv. foto's gegenereerd door generatieve AI) of gemanipuleerde informatie (bijv. foto's bewerkt met behulp van Photoshop), waardoor het voor gewone mensen (en zelfs voor professionals) moeilijk is om onderscheid te maken tussen echte en nep-/gemanipuleerde content. Bij het delen van foto's kunnen fotomakers, uitgevers en consumenten met behulp van data lineage erachter komen hoe en door wie een foto is gemaakt en welke bewerking(en) de foto heeft ondergaan gedurende de hele levenscyclus/ontwikkelingstraject. De herkomstinformatie, die betrouwbaar aan de foto kan worden toegevoegd en de foto tijdens zijn hele reis vergezelt, kan eenvoudig worden ingezien door consumenten stroomafwaarts in de pijplijn, zodat ze op de hoogte zijn van de herkomst en bewerkingsgeschiedenis van de foto. Deze kennis kan consumenten stroomafwaarts in de pijplijn helpen vertrouwen te krijgen in de content die ze tegenkomen op bijvoorbeeld sociale media en nieuwsfeeds. Binnen het Nederlandse rechtsstelsel kan data lineage op soortgelijke wijze bijdragen tot groter vertrouwen in bestaand beleid, bijvoorbeeld door vragen te beantwoorden als *welke datasets of documenten worden gebruikt als bewijs om het beleid te staven*.

In dit verslag beschrijven we de resultaten van ons verkennend onderzoek naar data lineage (en data provenance)-technologie, met name over de gedachte achter data lineage en de methoden/tools die gebruikt worden voor data lineage. De onderzoekscontext heeft betrekking op de data die worden verzameld, gedeeld, opgeslagen en verwerkt door informatiesystemen binnen het Nederlandse rechtsstelsel.

Doel van het onderzoek en onderzoeksvragen

Het doel van het onderzoek is erachter te komen hoe data lineage-technologie kan bijdragen aan data-governance en databeheer binnen het Nederlandse rechtsstelsel. Om dit doel te bereiken, moeten de voordelen van data lineage-technologie en de toepassingsmogelijkheden (en uitdagingen) binnen het Nederlandse rechtsstelsel worden onderzocht.

Dit is een vooronderzoek naar bovengenoemde doelstelling. De onderzoeksaanpak kan worden gekarakteriseerd als verkennend, waarbij we antwoorden zoeken op de volgende onderzoeksvragen:

1 *Wat is data lineage?* Om deze vraag te beantwoorden, zullen we ook de context (of het data-ecosysteem) waarin data lineage wordt gebruikt, beschrijven.
2 *Aan welke doelstellingen kan data lineage bijdragen?* Bij het beantwoorden van deze onderzoeksvraag gaan we in op de potentiële voordelen van data lineage.
3 *Hoe kunnen data lineage-tools worden ingezet?* Om antwoord te vinden op deze vraag, gaan we dieper in op de gebruikelijke manieren waarop data lineage wordt ingezet, en de uitdagingen die daarbij spelen.
4 *Wat zijn de mogelijkheden (en beperkingen) van bestaande data lineage-tools?* Om deze vraag te beantwoorden, schetsen we een kader voor het specificeren van de relevante data lineage-mogelijkheden. Voor een beperkt aantal bestaande data lineage-tools schetsen we twee toepassingsmogelijkheden die van belang zijn voor dit onderzoek.

Toepassingsgebied

Deze onderzoeksvragen zullen worden behandeld in het kader van het Nederlandse rechtsstelsel, dat bestaat uit veel semi-autonome organisaties die gezamenlijk de beginselen van de rechtsstaat in de Nederlandse samenleving waarborgen. De relaties tussen deze organisaties worden vaak gekenmerkt als een lineaire keten (waarbij een fase moet worden afgesloten voordat de volgende fase kan beginnen), met soms lussen en parallelle relaties. De term 'rechtsstelsel' verwijst naar de organisaties in het rechtsapparaat die betrokken zijn bij het produceren van data, variërend van wetteksten tot rechterlijke uitspraken. De werkingssfeer van het rechtsstelsel is ruimer dan dat van rechtbanken en gerechtelijke procedures.

In deze bijdrage willen we een overzicht geven van enkele belangrijke aspecten die kunnen worden overwogen bij het inzetten van data lineage in de organisatorische setting van het Nederlandse rechtsstelsel. Ons doel is niet om in deze bijdrage een oplossing te formuleren of voor te schrijven voor de implementatie van data lineage. De doelgroep van dit onderzoeksverslag zijn systeemontwerpers en -architecten, datafunctionarissen en dataspecialisten. Dit rapport heeft tot doel deze groepen te informeren over de ontwerpruimte waarbinnen zij een data lineage-oplossing kunnen ontwerpen of kiezen.

Methodologie

Voor dit onderzoek hebben we de literatuur kritisch bestudeerd, waarbij verschillende geselecteerde informatiebronnen zijn geanalyseerd, en hebben we nagedacht over de bestaande concepten, methoden en benaderingen. De geselecteerde informatiebronnen zijn niet alleen afkomstig uit wetenschappelijke literatuur, maar ook uit 'grijze' literatuur zoals commerciële websites, whitepapers en weblogs. Dit laatste

komt doordat veel leveranciers en systeemontwikkelaars voornamelijk actief zijn op het gebied van data lineage en veel innovatieve concepten, functies en tools in het domein introduceren.

Naast de literatuurstudie hebben we vier semi-gestructureerde interviews gehouden met deskundigen van verschillende organisaties binnen het Nederlandse rechtsstelsel om inzicht te krijgen in lopende data lineage-gerelateerde (R&D) activiteiten binnen het rechtsstelsel, en om de behoeften en opvattingen te peilen van de deskundigen die betrokken zijn bij data-governance/-management binnen hun organisatie. Verder hebben we twee focusgroepgesprekken georganiseerd met data-stewards en databeheer-experts om onze tussentijdse resultaten te presenteren en vroegtijdig feedback te krijgen. De vier geïnterviewden en de twee focusgroepen werden geselecteerd op basis van de deskundigheid en beschikbaarheid van de deelnemers en niet zozeer vanwege hun representativiteit. Deze keuze werd ingegeven door de voorlopige en verkennende aard van het onderzoek.

Belangrijkste resultaten en bijdragen

In dit verslag geven we een overzicht van verschillende aspecten van data lineage-technologie en hoe dit wordt ingezet in verschillende organisatorische omgevingen. Het verslag kan dienen als kennisbasis bij het ontwerpen en toepassen van data lineage in het Nederlandse rechtsstelsel. Hieronder volgt een samenvatting van de antwoorden op de gestelde onderzoeksvragen.

Wat is data lineage?

Op basis van enkele bestaande definities en de inzichten die tijdens het onderzoek naar data lineage zijn verkregen, definiëren we *data lineage* als volgt: *de beschrijving van databewegingen en -transformaties op verschillende abstractieniveaus langs datatrajecten*. *De beschrijving omvat de aspecten die van belang zijn in een toepassingscontext, zoals hoe (d.w.z. door wie, wanneer, waar, welke, enz.) dataobjecten worden verwerkt (d.w.z. gemaakt, verzameld, opgeslagen, geopend, getransformeerd, verzonden, enz.) en hoe deze gerelateerd zijn aan dataconcepten van hoog niveau.* Deze definitie conceptualiseert niet alleen de fysieke distributie van dataobjecten (zoals de oorsprong, stromen en transformaties van data), maar ook de semantische distributie van de gerelateerde concepten (zoals de business, juridische en organisatorische termen die betrekking hebben op, of van toepassing zijn op deze dataobjecten). Voorts biedt de definitie een middel om het toepassingsgebied van data lineage te beperken tot de aspecten die in elke context van belang zijn.

Data lineage draagt bij aan het vergroten van het vertrouwen in data en in verantwoorde datatransformatie en data-uitwisseling. Data lineage is afhankelijk van het afleiden en beheren van metadata die relevant zijn voor de beoogde gebruiksdoelstellling(en) van data lineage. Data lineage kent verschillende kenmerken, die niet noodzakelijkerwijs onafhankelijk van elkaar zijn. Deze kenmerken omvatten (a) oorsprong van data versus *lineage* van datastromen, (b) het 'waar' versus het 'hoe' van data lineage, (c) typen datatransformatie, (d) grofkorrelige versus fijnkorrelige data lineage, (e) luie versus gretige data lineage, (f) achterwaartse versus voorwaartse data lineage, (g) volgen versus traceren van data lineage, (h) technische versus business data lineage, en (i) horizontale versus verticale data lineage. Deze kenmerken van data lineage markeren de technische ruimte waarbinnen data lineage-oplossingen kunnen worden ontworpen, gekozen en/of ingezet.

Aan welke doelstellingen kan data lineage bijdragen?

Het *lineage*-concept is toegepast op een groot aantal gevallen, variërend van het volgen/traceren van de datastroomberekeningen in één softwareprogramma tot het traceren/traceren van datastromen in verspreide informatiesystemen. Data lineage kan bijdragen aan vele doelstellingen, die elk op hun beurt een rol spelen bij het vergroten van het vertrouwen in data, het delen van data, en datagestuurde applicaties en beleidsvorming. Deze doelstellingen omvatten a) data-governance, b) privacybescherming, c) vertrouwen in AI-modellen, d) data en AI uitlegbaarheid, interpreteerbaarheid en eerlijkheid, e) kwaliteitsbeheer van data, f) beheer van datawijzigingen, g) eigendom van data, h) naleving van regelgeving, controle en verantwoordingsplicht, i) databeveiliging, j) datamodellering en k) datadetectie. Deze doelstellingen beslaan de gebruiksgebieden van data lineage en specificeren als zodanig de *maatschappelijke relevantie van data lineage* in het algemeen.

De uitdaging is om bij de toepassing van een data lineage-oplossing alle genoemde doelstellingen in het vizier te hebben. Een dergelijke alomvattende data lineage-oplossing zou meteen te complex en kostbaar worden en dus niet realiseerbaar, met name in federatieve settings (bijv. in de semi-autonome organisaties van het Nederlandse rechtsstelsel). Als de doelstellingen voor data lineage duidelijk zijn, kan worden bepaald welke kenmerken van data lineage relevant zijn in een operationele setting. Op basis van de vereiste kenmerken van data lineage kan worden bepaald welk (type) metadata moet worden verzameld, en kan er vervolgens data lineage (implementatie)-architectuur worden ontworpen om metadata van data lineage op te slaan, te doorzoeken, te verwerken en op te halen (d.w.z. kan er een beslissing worden genomen over de architectuur voor het beheren van data lineage-metadata).

Hoe kunnen data lineage-tools worden ingezet?

Op basis van onze literatuurstudie en interviews met deskundigen concluderen we dat data lineage binnen het Nederlandse rechtsstelsel vooral kan (of moet) bijdragen aan data-governance, data discovery, kwailteitsbeheer van data, wijzigingsbeheer van data, en privacy en beveiliging. Verder zijn we tot de volgende conclusies gekomen over data lineage in het Nederlandse rechtsstelsel.

- Het is van belang dat gegevens binnen en tussen organisaties in het Nederlandse rechtsstelsel op betrouwbare wijze kunnen worden gedeeld, en dat deelnemende organisaties hun autonomie behouden en over eigen business, conceptuele en logische datamodellen en informatiesystemen beschikken. Data lineage binnen het Nederlandse rechtsstelsel moet rekening houden met diversiteit op alle niveaus, d.w.z. op technisch, logisch, conceptueel en business niveau.
- Binnen het Nederlandse rechtsstelsel worden data niet alleen voor onderzoek en strategische doeleinden gedeeld en verwerkt, maar ook voor operationele doeleinden. De gedeelde data kunnen zich dus op verschillende aggregatieniveaus bevinden, d.w.z. op groeps- en individueel niveau, wat betekent dat data lineage op grof- en fijnkorrelig niveau nodig is.
- De eindgebruikers van data lineage binnen het Nederlandse rechtsstelsel kunnen verschillende achtergronden en verschillende niveaus van datakennis- en -vaardigheden hebben, en bevinden zich mogelijk aan het begin, in het midden of aan het einde van de datapijplijn. Daarom moet er voldoende flexibiliteit zijn om een breed scala van gebruikers te bedienen en moet worden gestreefd naar het bredere bruikbaar maken van data lineage (zogenoemd de democratisering van data lineage).

- We definiëren vijf abstractieniveaus voor data lineage, te weten (a) fysiek, (b) logisch, (c) conceptueel, (d) business en (e) juridisch/ethisch. Het juridische en ethische aspect is bijzonder belangrijk in het Nederlandse rechtsstelsel, aangezien het stelsel verantwoordelijk is voor het toezicht op en de bescherming van de rechtsstaat.
- In organisatieoverschrijdende settings (zoals die van het Nederlandse rechtsstelsel) voorzien we dat horizontale data lineage niet alleen op fysiek niveau kan worden toegepast, maar ook op business en conceptueel niveau. Verder hebben we mogelijk een gecombineerde verticale en horizontale data lineage nodig om een horizontale data lineage op fysiek niveau te realiseren en zo interoperabiliteitsproblemen met betrekking tot data lineage aan te pakken (d.w.z. een verweven horizontale en verticale lineage-configuratie).

Het beheer van metadata in het Nederlandse rechtsstelsel moet schaalbaar en verspreid zijn en moet passen bij de organisatiestructuur van het stelsel. Gezien de organisatieoverschrijdende structuur van het rechtsstelsel, stellen we voor om een federatieve architectuur voor het beheer van data lineage-metadata te overwegen bij de inzet van data lineage. Een federatieve architectuur is gebaseerd op de bestaande organisatiestructuur, die daardoor organisch kan worden uitgebreid, zoals het geval is bij federatieve identiteitsbeheer tussen Europese universiteiten.

Voor het metadatabeheer van data lineage moet het proces voor het verzamelen van metadata worden geautomatiseerd, gezien de hoge snelheid waarmee data tegenwoordig worden verzameld en gedeeld, en het grote volume ervan. De keuze voor het type opslag van data lineage-metadata moet zijn gebaseerd op de vereiste data lineage-kenmerken en de context waarin data lineage wordt toegepast. Het type dataopslag bepaalt de keuze voor een geschikte verwerkingstaal voor zoekopdrachten. De interactie tussen systeem en gebruiker kan op verschillende manieren worden bevorderd, afhankelijk van de technische vaardigheden van de gebruikers.

Businessgerichtheid is een noodzaak, aangezien datagestuurd werken tegenwoordig de gangbare praktijk is geworden. Dit vereist dat niet-technische eindgebruikers bij het ontwerpproces moeten worden betrokken, en dat nieuwe zoekopdrachten automatisch kunnen worden opgenomen in data lineage-systemen. Het is echter niet haalbaar en efficiënt om uitgebreid een enorme hoeveelheid data lineage-metadata te verzamelen en te beheren, en dan af te wachten tot een deel van deze gegevens op een toekomstig moment bruikbaar is voor het beantwoorden van nieuwe business gestuurde zoekopdrachten. Om kostenefficiënter te zijn, kan ervoor worden gekozen een redelijke hoeveelheid data lineage-metadata te verzamelen (d.w.z. het 'grofkorrelig' verzamelen van metadata) en als er een nieuwe opzoekopdracht komt, kan men ofwel een gerichte zoekopdracht doen (d.w.z. een gefocuste zoekactie op need-to-know-basis) of een om-en-nabij antwoord geven als een zekere mate van onzekerheid in de antwoorden aanvaardbaar is. Wellicht is het nodig om op natuurlijke taal gebaseerde gebruikersinterfaces in te voeren die zowel vooraf gedefinieerde als niet vooraf gedefinieerde zoekopdrachten kunnen verwerken. Hiermee maak je data lineage business gestuurd en bruikbaar voor gebruikers met een beperkte technische achtergrond (zoals bedrijfsadviseurs en beleidsmakers).

Wat zijn de mogelijkheden (en beperkingen) van bestaande data lineage-tools?

Het is belangrijk de toepassingsmogelijkheden van data lineage-tools te bepalen in elke gebruikscontext. Ze kunnen namelijk worden gebruikt als criteria voor het

evalueren van de bestaande tools en bij de keuze welke tool het beste past bij de context. Op basis van de resultaten van het onderzoek voorzien we verschillende toepassingsmogelijkheden voor data lineage softwaretools in het Nederlandse rechtsstelsel. De belangrijkste mogelijkheden zijn: flexibiliteit om businessgericht te zijn (reageren op nieuwe en onvoorziene zoekopdrachten met betrekking tot data lineage), meer gedetailleerde lineage dan op attribuutniveau (zoals celniveau), interoperabiliteit met andere tools (van andere organisaties), en een goede gebruikerservaring kunnen bieden.

Als onderdeel van het antwoord op deze onderzoeksvraag schetsen we een kader voor het evalueren van data lineage-tools. Op basis van dit kader moet een reeks evaluatiecriteria worden gedefinieerd door bijvoorbeeld de doelstellingen te verduidelijken die justitiële organisaties nastreven bij het implementeren van data lineage, de gewenste data lineage-kenmerken te verduidelijken, en de evaluatiecriteria dienovereenkomstig te definiëren. Vervolgens kunnen per criterium verschillende niveaus van gedetailleerdheid worden gespecificeerd. Deze niveaus kunnen kwalitatief worden gedefinieerd, d.w.z. dat ze een (betekenisvolle) numerieke waarde of score toegekend krijgen. Ten slotte kan men gekwantificeerde criteria samenvoegen met behulp van numerieke methoden zoals middelen en het vaststellen van drempelwaarden.

Commerciële softwaretools bieden over het algemeen een breed scala aan data lineage-functionaliteiten, in combinatie met andere functionaliteiten (zoals databeheer en datacatalogus). Ze zijn geschikt voor grote organisaties die zich de kosten van dergelijke tools kunnen veroorloven en die gebruik moeten kunnen maken van het brede scala aan functies die deze tools bieden. Open-source- softwaretools bieden over het algemeen een beperkte subset aan data lineage-functionaliteiten, zijn gratis en kunnen worden geïntegreerd met andere (open-source-)applicaties. Daarom zijn ze geschikt voor kleine bedrijven met een gering budget. Zij kunnen deze tools gebruiken of ze aanpassen aan hun behoeften op het gebied van data lineage. Een uitgebreide integratie van open-source-tools vereist interne technische vaardigheden of extra middelen, die mogelijk niet beschikbaar zijn in kleine organisaties. Verder kunnen de open-source-tools nuttig zijn voor het uitvoeren van kleinschalige experimenten binnen de justitiële instelling om praktische ervaring op te doen met data lineage-technologie.

Aanbevelingen voor vervolgonderzoek

Tijdens het uitvoeren van projecten worden diverse toekomstige onderzoekspaden vastgesteld. In deze paragraaf groeperen we ze in drie categorieën, variërend van praktijkgericht onderzoek tot toegepast onderzoek.

Behoefte aan een onderzoek naar bestek van eisen

Bij het kiezen van een (set van) data lineage-tool(s) waarmee kan worden geëxperimenteerd (of die kunnen worden ingezet) in de justitiële instelling, erkennen we de noodzaak om de vereisten te verduidelijken waaraan de tool(s) moeten voldoen. Daarbij horen de volgende actiepunten:

- Bepalen van de typische eindgebruikers en hun data lineage-zoekopdrachten;
- nagaan of het nodig is de zoekopdrachten van de eindgebruikers in de toekomst aan te passen;
- definiëren van de gewenste doelstellingen/kenmerken van data lineage; en

- vaststellen van de evaluatiemethode voor de data lineage-tool door de evaluatiecriteria te bepalen, en hoe deze moeten worden gemeten en samengevoegd.

Als het doel is om data lineage in het Nederlandse rechtsstelsel te implementeren, moet verder onderzoek worden gedaan naar een geschikte structurele en functionele architectuur voor het inzetten van data lineage in het rechtsstelsel.

Ook zou onderzoek moeten worden gedaan naar de voorwaarden en manieren voor het mengen van verticale en horizontale datalijnen op de grenzen tussen organisaties. Daarvoor is het nodig dat de datasemantiek op de grenzen tussen samenwerkende organisaties in kaart wordt gebracht. Verder is de vraag hoe om te gaan met onzekerheden die kunnen worden veroorzaakt door deze inventarisatie van datasemantiek. Hiertoe kan het nuttig zijn om ervaring op te doen met data lineage-tools in de dagelijkse praktijk.

Behoefte aan het bredere bruikbaar maken van data lineage

Er is behoefte aan verder onderzoek naar het businessgericht en bruikbaar maken van data lineage voor gebruikers met een beperkte technische achtergrond (zoals bedrijfsadviseurs en beleidsmakers). Een veelbelovend onderzoeksgebied in dit verband zijn de methoden en tools voor op natuurlijke taal gebaseerde gebruikersinterfaces. Large Language Models (LLM's) kunnen worden ingezet voor het koppelen van teksten in natuurlijke taal aan formele database-zoekopdrachten (bijvoorbeeld SQL). Wellicht is het dan nodig te onderzoeken hoe met onzekerheid kan worden omgegaan bij de afstemming tussen natuurlijke en formele talen.

Behoefte aan effectieve en efficiënte data lineage

Voor een succesvolle implementatie van data lineage-technologie is het cruciaal om de complexiteit van data lineage en de bijbehorende kosten te verminderen.

Voor het verzamelen van data lineage-metadata is het noodzakelijk om geautomatiseerde methoden te ontwikkelen. Zo kan het gebruik van LLM's worden onderzocht voor het (semi-)automatisch genereren van metadata op business niveau (zoals een rapport in natuurlijke taal waarin de technische analyses worden beschreven die worden uitgevoerd op de gegevens voor eindgebruikers op business niveau) op basis van metadata op technisch niveau (bijvoorbeeld van dataquery-scripts). Op deze manier kan de belasting van het genereren van metadata op business niveau, voor technische deskundigen worden verlicht.

In conventionele data lineage wordt aangenomen dat gebruikers begrijpen dat een output wordt gegenereerd door het observeren van de brongegevens en dat zij weten dat de datatransformatie een reeks eenvoudige bewerkingen inhoudt, zoals filteren, samenvoegen en aggregeren. Bij complexe data-analyse, zoals het gebruik van AI/ML-algoritmen, is echter meer informatie over datatransformatie nodig. Een vraag die zich kan voordoen is welke data lineage-informatie moet worden verstrekt om de resultaten van zeer complexe datatransformaties (zoals LLM's) te verklaren en/of te beïnvloeden, en hoe deze data lineage-informatie op een (kostenefficiënte) manier kan worden beheerd.

Een probleem bij het bepalen van de herkomst van data is hoe de herkomst van data in elke setting moet worden bepaald. We vermoeden dat er meerdere opvattingen zijn met verschillende herkomsten van data (met name in organisatieoverschrijdende settings). Als deze hypothese standhoudt, dan vereist het bepalen van de herkomst van data het traceren van gegevensstromen. In toekomstig onderzoek moet worden gekeken naar hoe de herkomst van data (of bestemming van data) in operationele settings kan worden bepaald.

The WODC (Research and Data Centre), a Dutch agency in the field of Justice and Security, is an independent knowledge institute that falls under the Dutch Ministry of Justice and Security. The WODC contributes to upholding and improving the rule of law by carrying out high-quality scientific research (or commissioning others to do so on its behalf), as well as by presenting solicited and unsolicited knowledge, points for improvement and (where possible) thinking strategies.

More information:

www.wodc.nl