Datakwaliteit

Onderzoek naar indicatoren van datakwaliteit en de vertaling naar (automatische) toetsing

Spijker Mans | 03/02/2021

Onderzoek naar datakwaliteit

Auteur: Spijker Mans
School: HAS-hogeschool te ‘s-Hertogenbosch
In opdracht voor: Gunneman GIS & Geomatics
Versie: 1

Voorwoord

Allereerst wil ik Jochgem Gunneman bedanken voor het aanbieden van deze projectstage en de goede begeleiding gedurende de gehele stageperiode. Als tweede wil ik alle mensen bedanken waarmee ik persoonlijke gesprekken heb gehad, mede dankzij deze professionals en experts is het onderzoek van de grond gekomen en heeft het veel nieuwe inzichten opgeleverd. Als derde wil ik alle respondenten van de enquête bedanken voor het invullen. Door de respondenten heeft het onderzoek een betrouwbaarder resultaat. Tenslotte wil ik Marien de Bakker bedanken voor de begeleiding gedurende mijn stageperiode.

Samenvatting

De hoeveelheid data groeit met de seconde en veel organisaties hebben hun datakwaliteit onvoldoende onder controle. Datakwaliteit is een heel breed begrip en de toetsing van datakwaliteit is nog breder. Datakwaliteit speelt een grote rol in de uitkomst van gegevens of informatie. Data van mindere kwaliteit kan zorgen voor verkeerde voorspellingen of onjuiste informatie weergeven. Data van hogere kwaliteit kan geld besparen, goede informatie genereren en uitspraken gebaseerd op de data kunnen met meer zekerheid gedaan worden.

Het doel van dit onderzoek is om meer kennis en consensus te brengen bij bedrijven of organisaties die bezig (willen) zijn met datakwaliteit, en daarbij een helpende hand bieden door het opzetten van een raamwerk voor de (automatische) toetsing van datakwaliteitsindicatoren. Daarnaast is een ander doel van dit onderzoek om uit te zoeken of er draagvlak is voor het ontwikkelen van een Data Quality Control Platform door Gunneman GIS & Geomatics.

Binnen dit rapport is onderzoek gedaan naar het vertalen van indicatoren van datakwaliteit naar (automatische) toetsing van datakwaliteit. De hoofdvraag luidt dan ook als volgt: Hoe kunnen indicatoren van datakwaliteit vertaald worden naar (automatische) toetsing van datakwaliteit?

Het onderzoek is uitgevoerd door interviews af te nemen met diverse experts/professionals op het gebied van data en datakwaliteit. Daarnaast is er een enquête uitgezet die het onderzoek naar het draagvlak voor een Data Quality Control Platform ondersteunt.

Uit het onderzoek is gebleken dat datakwaliteit een lastig te definiëren begrip blijft, iedereen heeft een eigen interpretatie van datakwaliteit, en daarbij kunnen de doelen en behoeften bij datakwaliteit verschillen. De ideale definitie van datakwaliteit is de definitie die het bedrijf of de organisatie daar zelf aan geeft. Daarnaast is er op basis van de interviews en desk research een nieuw raamwerk opgezet voor de (automatische) toetsing van diverse datakwaliteitsindicatoren in FME. Uit de enquête is gebleken dat datakwaliteit zeer belangrijk is om inzicht in te krijgen, dat er wel degelijk behoefte is naar automatische toetsing van (automatische) toetsing van aspecten en dat een dashboard een van de beste manieren zou zijn om inzicht in datakwaliteit te weergeven.

De conclusie die uit dit onderzoek gehaald kan worden is dat het overgrote deel van de mensen last heeft van onvoldoende datakwaliteit en dat zich binnen hun werk uit doordat er veel tijd en geld verloren gaat aan de gevolgen van onvoldoende datakwaliteit. Daarnaast hoeven niet alle mogelijke datakwaliteitsaspecten meegenomen worden, alleen de belangrijkste. Het doel waar de data voor dient moet bekeken worden en op basis daarvan kan er aan de slag gegaan worden met de indicatoren.

In vervolgonderzoek kunnen de enquêteresultaten uitgebreider geanalyseerd worden door Chi^2-testen uit te voeren en de conclusies verder uit te werken. Daarnaast zou het combineren van diverse raamwerken en een verdieping in FME of andere software ook verder uitgewerkt kunnen worden in vervolgonderzoek.

Inhoudsopgave

Voorwoord 2

Samenvatting 3

Inleiding 5

Methode en materiaal 6

1. Uiteenlopende definities van datakwaliteit 7

2. Belangen en belanghebbenden bij datakwaliteit 9

3. Oorzaken en gevolgen van hoge en lage datakwaliteit 10

- Oorzaken van minder goede datakwaliteit 10
- Positieve gevolgen van betere datakwaliteit 11
- Negatieve gevolgen van slechte datakwaliteit 12
- Te hoge datakwaliteit 12

4. Indicatoren van datakwaliteit 14

- Het belang van datakwaliteitsindicatoren 14
- Datakwaliteitsindicatoren raamwerk 14
- Betekenis datakwaliteitsindicatoren 15
- Meetbaarheid van indicatoren 16

5. Verbeteren van datakwaliteit 17

6. Automatisch toetsbare indicatoren 19

- Verschil met eerdere raamwerken 20
  - De plaat van AAT 20
  - DAMA-DMBOK(2) 20
- Automatisch toetsbare indicatoren in FME 21

7. De kijk van experts op datakwaliteit, indicatoren van datakwaliteit en de automatische toetsing daarvan 23

8. Enquêteresultaten onderzoek datakwaliteit 25

- Algemene vragen 25
- Vragen datakwaliteit 27

Discussie 33

Conclusie 34

Verwijzingen 36

Inleiding

“95% van de organisaties heeft de data onvoldoende op orde”, aldus director Business IT Ramon van den Heuvel en senior consultant Luc Idzinga van Improven (Harmelink, 2016).

Het is onbekend hoe (andere) organisaties met het borgen van datakwaliteit omgaan en of er een gemeenschappelijke deler bestaat qua definities en toetsing van datakwaliteit. Het vermoeden bestaat dat er onvoldoende kennis is rondom datakwaliteit en de toetsing daarvan, zoals genoemd in de verwijzing van Harmelink. Datakwaliteit is een belangrijk begrip waar bedrijven en organisaties vermoedelijk niet van weten hoe groot het belang is en wat de oorzaken en gevolgen van een te lage datakwaliteit zijn. Het vermoeden dat dit een veelvoorkomend probleem is, is de aanleiding voor het onderzoek.

De doelstelling van dit onderzoek is om te kijken naar de mogelijkheden van (automatisch) toetsbare indicatoren in FME. Onder de doelstelling valt ook het informeren van een breed publiek over het belang van datakwaliteit, adviezen geven over hoe de datakwaliteit eventueel verbeterd kan worden, het creëren van een nieuw raamwerk voor meer mogelijkheden voor geïnteresseerden in datakwaliteit en meer gezamenlijke overeenstemmingen binnen de datakwaliteit wereld.

De hoofdvraag van het onderzoek luidt als volgt:
Hoe kunnen indicatoren van datakwaliteit vertaald worden naar (automatische) toetsing van datakwaliteit?

Hierbij zijn de volgende deelvragen opgesteld:

1. Welke uiteenlopende definities van datakwaliteit zijn er?
2. Wat zijn de belangen en wie zijn de belanghebbenden bij datakwaliteit?
3. Wat zijn de oorzaken en gevolgen van een hoge en lage datakwaliteit?
4. Welke indicatoren zijn er voor datakwaliteit?
5. Hoe kan datakwaliteit verbeterd worden?
6. Welke indicatoren van datakwaliteit zijn automatisch toetsbaar?
7. Hoe kijken experts aan tegen datakwaliteit, indicatoren van datakwaliteit en de automatische toetsing daarvan?
8. Wat zijn de resultaten uit de enquête?

Methode en materiaal

Dit onderzoek is uitgevoerd door het gebruik van zowel kwalitatief als kwantitatief onderzoek. Het kwalitatieve onderzoek is uitgevoerd door het houden van interviews met mensen binnen de (geo-)data wereld. Via LinkedIn zijn deze experts en professionals benaderd met de vraag of er eventueel tijd was voor een kort gesprekje. De gesprekken waren veelal spontaan, maar af en toe de goede richting in gestuurd met gerichte vragen waar antwoord op werd gezocht voor het onderzoek. Het kwantitatieve onderzoek is uitgevoerd door middel van het uitzetten van een enquête. De enquêtetool die gebruikt is voor het onderzoek heet 'SurveyPlanet'. Het uitzetten van de enquête ging veelal via LinkedIn, maar ook persoonlijke benadering via e-mail. Daarnaast heeft 'Dux-Soup', een LinkedIn automatiseringsprogramma, ook een rol gespeeld in het werven van respondenten.

Voor het ontwikkelen van een voorzet voor automatisch toetsbare indicatoren heb ik de software FME gebruikt. FME, ofwel Feature Manipulation Engine, is een tool die data integratie, verwerking en ontsluiting kan automatiseren. De software biedt een automatisering van processen waarbij het gehele proces reproduceerbaar is.

1. Uiteenlopende definities van datakwaliteit

Om datakwaliteit te definiëren kan het woord eerst gesplitst worden in de termen ‘data’ en ‘kwaliteit’. Data is een verzamelnaam voor verschillende concepten zoals records, attributen, gegevenswaarden en metadata (Black & van Nederpelt, 2020). Een andere definitie voor data is simpelweg; ‘gegevens’. De kwaliteit van iets is de mate waarin datgene goed is of aan bepaalde normen voldoet (Ensie, 2011).

Datakwaliteit is een begrip waarvan de betekenis per branche, bedrijf, organisatie en zelfs persoon kan verschillen. Voor datakwaliteit geldt niet één goed antwoord of een bepaalde standaard die men kan hanteren. Tussen branches zul je vooral veel verschil horen in de definitie van datakwaliteit. Zo heeft de zorg bijvoorbeeld andere behoeftes bij datakwaliteit dan de agrarische sector.

“Fit for purpose” ofwel “Fit for use”, naar het Nederlands vertaald “doeltreffendheid”, is de definitie die veruit het meest voorkwam uit de interviews en op het internet. De betekenis van het begrip betekent het volgende: de mate waarin de gegevens geschikt zijn voor het doel waarvoor je ze wilt gebruiken.
Fit for purpose is een definitie gegeven door Richard Y. Wang en Diane M. Strong in het boek ‘What Data Quality Means to Data Consumers (1996)’.
De betekenis is breed te interpreteren en zal voor velen tot andere uitkomsten leiden. Als er een doel is opgesteld waarvoor je de gegevens wilt gebruiken, blijven er nog steeds de vragen; wat is geschikt en wanneer is data geschikt?

Een ander veelvoorkomende definitie die datakwaliteit kan omschrijven is ‘NAVJ’. Dit is een afkorting voor vier indicatoren die gebruikt worden om datakwaliteit te toetsen, namelijk:

- Nauwkeurigheid;
- Actualiteit;
- Volledigheid;
- Juistheid.

Dit is een concrete definitie die een aantal aspecten van datakwaliteit goed naar voren brengt. Met de NAVJ definitie worden een paar belangrijke eisen voor het toetsen van datakwaliteit benoemd, maar er zijn nog een hoop andere mogelijke indicatoren die nu niet meegenomen worden. NAVJ geeft wel een duidelijker beeld hoe er getoetst kan worden op datakwaliteit. Bij de definitie NAVJ geldt ook, het is afhankelijk van het doel waarvoor je de data gebruikt. Meer over indicatoren in hoofdstuk 4, ‘indicatoren van datakwaliteit’.

Een definitie van datakwaliteit van de website biplatform luidt als volgt: “Het is belangrijk dat iedereen die waarde hecht aan de inhoud van de gegevens hetzelfde bedoelt” (Klaassen, 2004). Dit sluit aan op het ‘fit for purpose’ principe, maar het definieert datakwaliteit nog sterker. Deze definitie betrekt ook de menselijke kant van datakwaliteit, wat veelzijdigheid aan het begrip toevoegt.

In een interview met Harl van Heertum (Heertum, 2020) werd het volgende genoemd: “Er zijn eigenlijk 2 definities voor datakwaliteit. De ene definitie focust zich op de kwaliteit van de data zelf, de andere definitie focust zich op het proces van de datakwaliteit.” Dit is een interessante benadering van het begrip omdat het een tweedeling maakt voor verschillende dimensies. Door een tweedeling te maken wordt het begrip specifieker en makkelijker te implementeren en controleren.

De laatste definitie van datakwaliteit kwam uit een interview met R. Voorter: “De kwaliteit van de data die je uit de database kan halen moet hetzelfde zijn als de werkelijke data waarmee de mensen in het veld aan de slag gaan” (Voorter, 2020). Hierbij werd aangegeven dat deze opvatting vooral gebaseerd is op hoe het in de praktijk ervaren wordt en niet zozeer gebaseerd is op onderzoek.

Al met al kan er gesteld worden dat er geen eenduidige definitie van het begrip datakwaliteit bestaat. Wel kan er uit de verschillende definities die besproken zijn in dit hoofdstuk worden gesteld dat het belangrijk is dat er gekeken wordt naar het desbetreffende bedrijf of de organisatie en de data waar mee gewerkt wordt. Er moet een onderbouwde keuze gemaakt worden in de definitie voor datakwaliteit die aansluit bij de wensen van een organisatie of bedrijf.

Welke definitie er ook gekozen wordt, er kunnen allerlei vragen opborrelen als;

- Wat is geschikt?
- Hoeveel afwijking is acceptabel?
- Hoe weet ik dat mijn gegevens de waarheid voorstellen?
- Waarom überhaupt toetsen op datakwaliteit?
- Wie hebben er belang bij datakwaliteit?

Het antwoord op de laatste vraag wordt in het volgende hoofdstuk bekeken.

2. Belangen en belanghebbenden bij datakwaliteit

Erkennen van het belang van datakwaliteit is stap één naar een verbeterde datakwaliteit. Wanneer er geen erkenning vanuit management en/of personeel is zullen verdere stappen weinig tot geen vruchten afwerpen.

Datamanagement begint met het besef dat data belangrijk is voor de bedrijfsvoering. Data awareness bij het management is een randvoorwaarde. Data awareness bij de medewerkers is noodzakelijk om het door te voeren in de dagelijkse praktijk (Conijn, De Jong, & Krol, 2019).

De concrete belangen bij datakwaliteit zijn tijd, geld en inspanning. IBM (International Business Machines Corporation) schatte de jaarlijkse kosten van slechte datakwaliteit in de Verenigde Staten alleen al op 3,1 biljoen dollar (Redman, 2016). Wat tijd, geld en inspanning verder inhoudt wordt verderop uitgelegd in hoofdstuk 3 onder ‘positieve gevolgen van betere datakwaliteit’.

Iedereen heeft direct of indirect baat bij een goede datakwaliteit. Het is belangrijk dat zoveel mogelijk mensen die betrekking hebben tot de data de meerwaarde inzien van een verbeterde datakwaliteit. Mensen die direct betrekking hebben tot de data zijn bijvoorbeeld; data-analisten, data-scientists, data-visualisers en GIS specialisten. Ook de mensen die niet direct met data werken maar wel over besluitvorming of advisering gaan op basis van die data zoals consultants en beleidsmakers moeten de meerwaarde inzien. Als laatste is het goed voor eindgebruikers en/of klanten om inzicht te krijgen in de kwaliteit van de aangeleverde data.

Het belang van datakwaliteit wordt verder inzichtelijk wanneer de positieve gevolgen van een betere datakwaliteit en negatieve gevolgen van een slechte datakwaliteit duidelijk zijn. Dit wordt nu verder behandeld in hoofdstuk 3.

3. Oorzaken en gevolgen van hoge en lage datakwaliteit

Oorzaken van minder goede datakwaliteit

Er zijn een aantal oorzaken te noemen die voor een lagere datakwaliteit kunnen zorgen. De genoemde oorzaken staan niet in specifieke volgorde.

De eerste oorzaak van een lagere datakwaliteit ligt al bij de onnauwkeurigheid van de notitie van bronnen (Klink, 2020). Niks kan gecontroleerd worden als de bron van de data niet duidelijk is. Dit valt onder een stukje metadata die niet goed is bijgehouden. Weinig, geen of slecht bijgehouden metadata haalt de betrouwbaarheid van de data al sterk omlaag.

De tweede oorzaak die te noemen is voor een lagere datakwaliteit heeft te maken met de inwinning van de data. Afwijkingen in meetapparatuur of verkeerd gebruik van meetapparatuur kunnen voor een (te) lage nauwkeurigheid zorgen. Afwijkingen in meetapparatuur moeten op voorhand al meegenomen worden bij het invoeren van de data óf duidelijk beschreven staan in de metadata zodat de afwijking bij verdere berekeningen meegenomen kan worden.

Als derde is onvoldoende standaardisatie een enorm belangrijke oorzaak van lagere datakwaliteit. Dat standaardisatie belangrijk is blijkt uit de gehouden interviews. Standaardisatie werd door velen benoemd als belangrijke oorzaak van lage datakwaliteit. Standaardisatie houdt in dat data gerapporteerd wordt op dezelfde manier. Dit kan benaming, aantal decimalen, datumnotatie en nog veel meer zijn. Een aantal voordelen hierbij zijn; het (makkelijker) koppelen van gegevens aan elkaar, minder tijd kwijt met opschonen en de efficiëntie dat iedereen hetzelfde bedoeld. "*De grote uitdaging met standaardisatie is het overschakelen van de oude versie naar een nieuwe versie. Hierbij kan een hoop verloren gaan en het kan ook zorgen voor problemen bij werknemers. Nieuw beginnen is makkelijk, maar de historie overzetten is moeilijk*" (Kaper, 2020).

Als vierde oorzaak wordt in een interview met E. Poppe genoemd dat verkeerde doorberekeningen, bijvoorbeeld bij attributen die zijn afgeleid van andere attributen, de kwaliteit van de data kan verslechteren. Dit kan gebeuren wanneer data wordt getransformeerd, oftewel, wanneer nieuwe tabellen worden aangemaakt op basis van een berekening van andere tabellen. Als de data in de eerste instantie kwalitatief niet in orde is, zullen doorberekeningen gebaseerd zijn op verkeerde data.

De vijfde oorzaak voor minder goede datakwaliteit kwam ook naar voren in het interview met E. Poppe. Er werd benoemd dat onvoldoende gevoel tussen de relatie van gegevens en de werkelijkheid kan zorgen voor onnodige fouten. Wanneer waarden niet kloppen en iemand met verstand van de werkelijkheid ernaar kijkt, kunnen afwijkingen herkend worden. Hierbij kan ook een vier-ogen principe gehanteerd worden waarbij er altijd door twee personen gecontroleerd wordt of alles lijkt te kloppen.

Als laatste oorzaak voor slechtere datakwaliteit is er data niet aan het systeem toegevoegd wordt. Dit kan zijn omdat het belang van de data nog niet wordt ingezien. Dit gebeurt soms ten onrechte waardoor nuttige data niet gebruikt wordt. Andersom kan exact hetzelfde gebeuren, namelijk dat data die niet van meerwaarde is, of niet aan gestelde eisen voldoet, wél wordt meegenomen. Dit zorgt voor meer bezette opslagruimte, soms voor extra rekenkracht bij het uitvoeren van diverse analyses en kan een dataset/database onoverzichtelijker maken. Uiteraard zal het ook de kwaliteit van de data omlaag halen wanneer er getoetst wordt.

De hiervoor genoemde oorzaken hebben direct betrekking op de data. Naast de directe oorzaken zijn er ook oorzaken voor het niveau van de datakwaliteit die 'hoger' liggen. Deze hogere oorzaken zijn:

- Kwaliteit van de data governance (gegevensbeheer);
- Kwaliteit van de processen;
- Kwaliteit van de systemen;
- Kwaliteit van het informatie model;

De kwaliteit van bovenstaande oorzaken spelen een belangrijke rol in de uiteindelijke kwaliteit van de data. Binnen dit rapport wordt een uitleg over de oorzaken gegeven met een korte toelichting hoe dit invloed heeft op de datakwaliteit.

"Data governance is de specificatie van beslissingsrechten en een verantwoordingsraamwerk om het juiste gedrag te verzekeren bij de waardering, creatie, opslag, gebruik en controle van gegevens en analytics" (Gartner, 2021).
Data governance is een breed begrip met een hoop taken en verantwoordelijkheden. Dit is niet iets wat binnen een bedrijf of organisatie van de ene op de andere dag verschijnt. Indien er wel wordt gewerkt aan de data governance zal het helpen om juiste keuzes te maken rondom data wat een positieve invloed zal hebben op de datakwaliteit.

Kwaliteit van de processen moet in orde zijn om een overzicht te behouden binnen je data. Binnen de processen van een bedrijf of organisatie moet voorkomen worden dat er een brei aan applicaties gekenmerkt door een vervlechting van grote hoeveelheden interfaces ontstaat. Deze brei aan applicaties en interfaces kan ervoor zorgen dat processen ingewikkelder, en hiermee onoverzichtelijker, worden en meer tijd in beslag nemen.

De kwaliteit van de systemen speelt een belangrijke rol in de verwerking en opslag van gegevens. Wanneer de systemen bijvoorbeeld verouderd zijn of niet efficiënt functioneren kan het voorkomen dat je data daar onder lijdt. Verouderde systemen kunnen moeite hebben met de groeiende hoeveelheden data of nieuwe formats niet meer ondersteunen.

Een informatiemodel bestaat uit een grafische weergave van alle objecten en hun onderlinge verhouding en een tekstuele toelichten daarop, die zowel definities bevat als de bedrijfsregels die erop van toepassing zijn (Infomedics, 2017). Een informatiemodel kan een deelbare, stabiele en georganiseerde structuur van informatie-eisen of kennis bieden. Dit zorgt voor minder fouten rondom definities, relaties en regels.

Positieve gevolgen van betere datakwaliteit

Er zijn een hoop voordelen te noemen die komen kijken bij een verbeterde datakwaliteit.

Ten eerste zorgt verbeterde datakwaliteit voor snellere en betere besluitvorming. Betere gegevens leiden namelijk tot minder discussie en meer betrouwbare besluiten. Dit geeft meer vertrouwen in de besluiten van het management onder medewerkers.

Ten tweede zorgt verbeterde datakwaliteit voor meer efficiënte en voorspelbare processen. Betere gegevens voorkomen inefficiënte handelingen en onverwachte verrassingen. Goede datakwaliteit leidt tot besparingen op bijvoorbeeld inkoop- en productiekosten, onder meer vanwege het first-time-rightprincipe. Als alles op orde is krijg je later in het proces minder correctieslagen die geld kosten (Harmelink, 2016).

Ten derde zorgt verbeterde datakwaliteit voor compliance. Door betere gegevens kun je beter, efficiënter en met meer vertrouwen rapporteren over het voldoen aan wet- en regelgeving. Indirect zal dit ook onverwachtse kosten schelen zoals bijvoorbeeld boetes door overtreding van de wetten als de AVG.

Als laatste zorgt verbeterde datakwaliteit voor een eventueel concurrentievoordeel door business, financiële, niet-financiële en externe data te combineren en te gebruiken. Dit levert beter inzicht in je klanten en hun behoeften, maar bijvoorbeeld ook opkomende trends waardoor je sneller marktkansen waar kunt nemen (Weber, 2020).

Al de genoemde positieve gevolgen van verbeterde datakwaliteit zorgen voor meer tevreden werknemers, managers en klanten. Het geeft een goed beeld aan het bedrijf en klachten door triviale fouten zullen sterk verminderen. Naast de positieve gevolgen van een hoge datakwaliteit zijn er ook negatieve gevolgen verbonden aan een minder goede datakwaliteit.

Negatieve gevolgen van slechte datakwaliteit

Alle positieve gevolgen van een betere datakwaliteit kunnen omgedraaid worden naar negatieve gevolgen van minder goede datakwaliteit, maar er zijn ook losstaande negatieve gevolgen die opspelen bij een lagere datakwaliteit.

Allereerst zorgt het nemen van beslissingen op basis van foutieve data onnodige risico's met zich mee. Hoe hoger het risico op afwijkingen, des te hoger de kans op onverwachte kosten. Uit een interview met A. Klink kwam naar voren dat kleine afwijkingen in de data bij beheer van ondergrondse infra kabels, leidingen en riolering kunnen leiden tot grote problemen.

Als tweede zal het onder medewerkers zorgen voor een lager werkniveau en daarmee eventuele frustratie die daarbij komt kijken. De werknemers die zich bijvoorbeeld normaal bezig houden met analyse zullen nu veel werk hebben aan het oplossen en opschonen van triviale fouten (Bouman, 2005).

Als laatste kan een lage datakwaliteit zorgen voor een verslechterd bedrijfsimago bij stakeholders. Dit sluit aan op het eerst genoemde punt over het risico op afwijkingen. Neem hetzelfde voorbeeld van de ondergrondse infrastructuur, wanneer daar een fout op centimeters wordt gemaakt kan dat voor bewoners en gemeenten grote invloed hebben.

Te hoge datakwaliteit

Iedereen hecht andere waarden aan datakwaliteit en het nodige niveau van de datakwaliteit om de gestelde doelen te behalen. Een goede datakwaliteit is belangrijk zoals hierboven uitgebreid toegelicht is, maar is er ook een té hoge datakwaliteit en hoe kan dat zich voordoen?

Te hoge datakwaliteit houdt in dat de tijd en/of kosten van het verbeteren van de datakwaliteit hoger zijn dan de baten. Deze betekenis is binnen dit rapport gedefinieerd omdat hier (nog) maar weinig andere bronnen over gerapporteerd hebben.

Een te hoge datakwaliteit kan veroorzaakt worden door onvoldoende inzicht in de waarde van de data. Dit kan op een aantal manieren waaronder; data kwalitatief verbeteren die niet gebruikt wordt of te weinig oplevert, maar ook (dure) systemen implementeren die hun geld niet terugverdienen.

Het gevolg van een te hoge datakwaliteit is dat bedrijven of organisaties onnodig hoge kosten maken die op een gegeven moment niet meer opwegen tegen de verdiensten. Datakwaliteit verbeteren kan eenmaal duur zijn en wanneer data kwalitatief wordt verbeterd die niet van belang is wordt er onnodig geld uitgegeven.

4. Indicatoren van datakwaliteit

Datakwaliteit is abstract, kwaliteitsindicatoren maken het concreet (Conijn, De Jong, & Krol, 2019).

In hoofdstuk 1 “Uiteenlopende definities van datakwaliteit” werd duidelijk dat datakwaliteit een vrij abstract begrip is en dat er door verschillende mensen verschillende definities gegeven worden. Met de hulp van indicatoren voor datakwaliteit, vanaf nu ‘datakwaliteitsindicatoren’, wordt het begrip datakwaliteit concreter.

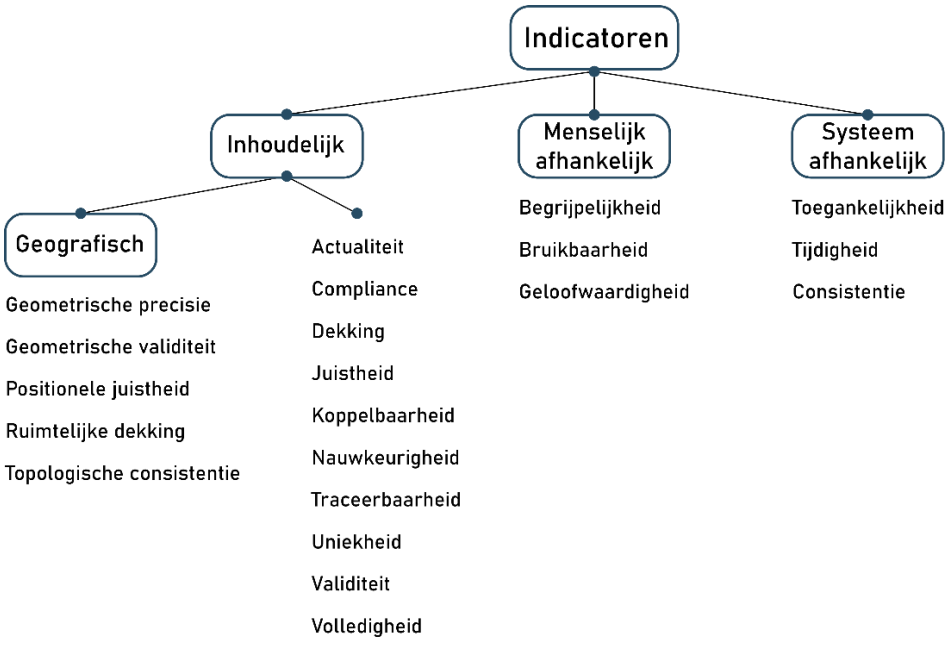
Het belang van datakwaliteitsindicatoren

Datakwaliteitsindicatoren moeten ervoor zorgen dat datakwaliteit meetbaar wordt. Welke indicatoren van belang zijn kan verschillen voor ieder doel waarvoor de data wordt gebruikt. In de volgende afbeelding is een overzicht te zien van verschillende indicatoren en waar deze onderverdeeld kunnen worden.

Datakwaliteitsindicatoren raamwerk

De volgende indicatoren komen vanuit diverse bronnen. De voornaamste bron is een eerder raamwerk genaamd, ‘De plaat van AAT’. Daarnaast komen er ook een aantal indicatoren en betekenissen uit het document ‘Gegevenskwaliteit in de Omgevingswet’.

Onderstaand raamwerk is gecreëerd voor dit onderzoek.



Figuur 1 Indicatoren datakwaliteit

Een korte toelichting op de onderverdeling:

Systeem afhankelijk houdt in dat de indicatoren voor toetsing afhankelijk zijn van het systeem wat gebruikt wordt. Om een voorbeeld te noemen, tijdigheid geeft aan in hoeverre de data op tijd beschikbaar is. Hiermee wordt bedoeld hoe lang het duurt voordat de data die uit het systeem wordt genomen beschikbaar is. Hoe lang het precies duurt hangt af van het systeem.

Menselijk afhankelijk houdt in dat de indicatoren voor toetsing afhankelijk zijn van menselijke interpretatie. Begrijpelijkheid is niet door een computer te definiëren en heeft menselijke verificatie nodig om de indicator te toetsen.

Geografische indicatoren zijn alle indicatoren die verwijzen naar gegevens en informatie die expliciet of impliciet geassocieerd zijn met een locatie (Omnicsci, sd).

De overige inhoudelijke indicatoren konden niet specifiek ergens onderverdeeld worden. Het zijn indicatoren die inhoudelijk betrekking hebben op de data.

Betekenis datakwaliteitsindicatoren

De bovenstaande indicatoren zijn eigenlijk nog vrij abstract, vooral als er nog weinig beeld is bij datakwaliteit. In de volgende tabel worden alle indicatoren genoemd met bijpassende betekenis. De definities kunnen verschillen tussen bronnen, zo heeft iedereen zijn eigen interpretatie. Binnen dit rapport worden de volgende definities gehanteerd:

Actualiteit	De mate waarin de gegevens actueel zijn.
Begrijpelijkheid	De mate waarin gegevens eenvoudig gelezen en geïnterpreteerd kunnen worden door gebruikers.
Bruikbaarheid/ toepasbaarheid	De mate waarin de registratie kan voldoen aan de gebruikerswensen voor een bepaalde toepassing.
Compliance	De mate waarin de gegevens in overeenstemming zijn met de geldende wet- en regelgeving.
Consistentie	In hoeverre alle waarden op dezelfde manier zijn bepaald en of door het hele systeem dezelfde codering en verwijzingen gehanteerd worden.
Dekking	De mate waarin de data een goede weergave is van de populatie (statistisch te onderbouwen)
Geloofwaardigheid/plausibiliteit	De mate waarin gegevens worden beschouwd als waar en geloofwaardig door gebruikers.
Geometrische precisie	De mate van detail waarmee ruimtelijke gegevens worden ingewonnen.
Geometrische validiteit	De mate waarin iedere geometrie voldoet aan geometrische regels.
Integriteit	Synoniem betrouwbaarheid. Beschermen van juistheid en consistentie
Juistheid	De mate waarin gegevens overeen komen met de werkelijkheid.
Koppelbaarheid	De mate van gemakkelijkheid waarmee datasets gekoppeld kunnen worden
Nauwkeurigheid	De mate van de nauwkeurigheid van gegevens. (bijv. aantal decimalen)
Positionele juistheid	De mate waarin locatiegegevens overeenkomen met de werkelijkheid.
Ruimtelijke dekking	De mate waarin de gegevens ruimtelijk dekkend zijn. Zijn er bijvoorbeeld ruimtes tussen polygonen waar dat niet hoort.
Tijdigheid	Geeft aan in hoeverre de data op tijd beschikbaar is
Toegankelijkheid	De mate van toegankelijk voor iedereen die er bij moet kunnen.
Topologische consistentie	De mate waarin ruimtelijke gegevens zich op de juiste wijze tot elkaar verhouden.
Traceerbaarheid	De mate waarin de totstandkoming en het gebruik van gegevens zijn vastgelegd.
Uniekheid	De mate waarin de waarden die niet dubbel voor mogen komen maar één keer voorkomen.
Validiteit	De mate waarin gegevens voldoen aan de verwachte structuur en opslagvorm.
Volledigheid	Geeft aan in hoeverre alle noodzakelijke data aanwezig is

Binnen een organisatie of bedrijf moet voor iedereen duidelijk zijn wat de indicatoren inhouden. Welke definitie er ook gebruikt wordt, iedereen moet hetzelfde bedoelen. Als dat niet het geval is kunnen begrippen voor onderlinge verwarring zorgen en maken indicatoren het niet concreter, maar juist onduidelijker.

Welke indicatoren precies relevant zijn voor het bedrijf of de organisatie moet per geval uitgezocht worden. Het is voornamelijk belangrijk dat indicatoren worden gekozen waar waarde aan gehecht wordt en waar de kosten opwegen tegen de baten, zoals toegelicht onder 'te hoge datakwaliteit'.

Meetbaarheid van indicatoren

Nu de indicatoren gedefinieerd zijn geeft dat een duidelijker beeld van wát er getoetst kan worden. De meetbaarheid van indicatoren is van belang om datakwaliteit concreet te kunnen toetsen. Er wordt toegelicht hoe de indicatoren meetbaar worden voor toetsing.

Om te beginnen kan het helpen om KPI's aan de indicatoren te hangen. Deze KPI's geven concrete waarden aan de indicatoren waardoor er doelgerichter gewerkt kan worden.

Een voorbeeld hiervan is met de indicator uniekheid (de mate waarin de waarden die niet dubbel voor mogen komen maar één keer voorkomen). Deze indicator kan een relatieve waarde krijgen, zoals 90%, om als meetlat te gebruiken. Wanneer het percentage uniekheid van velden lager ligt dan deze opgestelde waarde kan de indicator worden bestempeld met 'onvoldoende'. De KPI's kunnen ook minimale en maximale waarden zijn die toegestaan zijn binnen een bepaald veld. Om een ander voorbeeld te noemen; de oppervlakte van een vlak mag niet groter zijn dan x m^2 en het vlak mag ook niet kleiner zijn dan x m^2. Deze waarden kunnen op basis van de werkelijkheid opgesteld worden, maar ook op basis van preferentie om een bepaald doel na te streven. De waarden kunnen zoals in de voorbeelden genoemd zowel relatief als absoluut zijn.

Het stellen van KPI's kan goed werken bij indicatoren waar daadwerkelijk een (grens)waarde aan gegeven kan worden. Er zijn echter ook indicatoren die afhankelijk zijn van menselijke interpretatie zoals begrijpelijkheid. Om menselijk afhankelijke indicatoren meetbaar te maken zijn altijd meerdere personen nodig. Wat mensen begrijpelijk vinden is relatief, maak het concreter door vergelijkingen te maken. Vinden 3 op de 10 mensen het begrijpelijk, dan geeft dat door middel van concrete getallen aan dat er verbetering nodig is. Door in getallen uit te drukken worden menselijk afhankelijke indicatoren meetbaarder.

Een andere manier van controle op datakwaliteit is een controle met outlier modellen: Zijn er velden waartussen een verhouding zit, dan kun je op basis daarvan een inschatting maken of de waarden kloppen. Er kan dus een correlatie tussen velden zitten die zorgt voor een controle (Heertum, 2020).

5. Verbeteren van datakwaliteit

De oorzaken, gevolgen en indicatoren van datakwaliteit zijn inzichtelijk gemaakt in voorgaande hoofdstukken. Nu is het belangrijk om deze kennis daadwerkelijk toe te passen. In dit hoofdstuk wordt uitgebreid uitgelegd en onderbouwd hoe een bedrijf of organisatie concreet de datakwaliteit kan verbeteren.

Voor het aannemen van verschillende adviezen en het doorvoeren van nieuwe ideeën is het belangrijk om, zoals genoemd in hoofdstuk 2, in ieder geval voor de gehele organisatie duidelijk te hebben waarom de datakwaliteit wordt verbeterd. Bij het management is doordringen van het belang van datakwaliteit erg belangrijk, het management is uiteindelijk degene die beslissingen maakt op basis van de data.

Er volgen een aantal adviezen voor het verbeteren van datakwaliteit. De geschikte oplossingen en adviezen kunnen verschillen per doel dat wordt gesteld voor de data. De adviezen zijn niet in specifieke volgorde.

Als eerste moet voor iedereen helder zijn dat datakwaliteit een constant lopend project is. Datakwaliteit is geen tijdelijk project dat van de ene op de andere dag voltooid is, maar juist iets wat constant aandacht nodig heeft om continu te verbeteren of niet te laten verslechteren.

Daarnaast moet helder zijn wie het eigenaarschap draagt voor datakwaliteit. Is dit een heel team of is een enkele persoon verantwoordelijk? De business moet in de lead zijn. Niet de IT-afdeling. Die speelt vooral een faciliterende rol (Sogeti, sd).

Het is belangrijk dat alleen relevante data vastgelegd wordt. Vooral met buzzwoorden als ‘big data’ en ‘data gedreven werken’ kan het verleidelijk zijn om zoveel mogelijk data in te winnen. Zorg voor een goede structuur en win alleen data in die relevant is voor de organisatie. Om erachter te komen wat relevante data is moet het doel van het gebruik van de data duidelijk zijn.

Het kiezen van de juiste indicatoren en het stellen prioriteiten. De juiste indicatoren zijn voor iedereen verschillend. Kies indicatoren die aansluiten op het doel of de wensen. Stel prioriteiten wanneer het doel te groot is of er te veel wensen zijn, vooral wanneer toetsing niet meer rendabel is.

Technisch afdwingen van de invoer van data kan ook helpen om de datakwaliteit te verbeteren. Veel vrijheid laten aan de kant van invoer leidt tot vervuilde en/of onvolledige data. Dwing daarom technisch af dat data zo veel mogelijk hetzelfde wordt ingevoerd (Legerstee, 2019).

Zoals eerder genoemd in de oorzaken van een lagere datakwaliteit is eenheid en standaardisatie een van de grootste problemen bij bedrijven en organisaties. Om dit te verbeteren kunnen raamwerken ingezet worden. Diverse raamwerken worden verderop in dit onderzoek genoemd in Hoofdstuk 6 ‘Automatisch toetsbare indicatoren’. Oneliners zijn daarentegen ook een goed middel om in te zetten om definities van de aspecten van datakwaliteit voor iedereen eenduidig te maken. Dit voorkomt onderlinge verwarring binnen een team (Jutte, 2020).

Het verzamelen van data vindt niet alleen intern plaats. Data kan ook afkomstig zijn van bijvoorbeeld externe leveranciers. Om hoge datakwaliteit te -realiseren, is het belangrijk om standaarden voor uitwisseling af te spreken over formaten, volumes of tijdsperiodes (Legerstee, 2019).

Transparantie, zowel naar klanten als medewerkers, zorgt voor een groter vertrouwen in het bedrijf of de organisatie. Welke klantdata sla je op en welke data deel je met derden? Een bekend voorbeeld waar dat niet altijd transparant is, is facebook. Zij lopen imagoschade op door

onduidelijkheid rondom gegevens. Met datakwaliteit zal dit niet zo snel voorkomen, tenzij er fouten ontstaan door slechte datakwaliteit. Transparantie is daarom erg belangrijk om eventuele fouten te voorkomen.

Misschien wel een van de belangrijkste punten voor het verbeteren van datakwaliteit is monitoring. Door middel van een nulmeting kunnen vergelijkingen gemaakt worden met nieuwe metingen. Zorg voor standaardisatie bij het toetsen van datakwaliteit om zo concreet mogelijke vergelijkingen te maken. Trendanalyse per indicator zorgt voor duidelijk inzicht in het verloop van je datakwaliteit door de tijd heen. Door specifiek per indicator te kijken kan er ook specifieker gehandeld worden op de juiste indicator.

Aansluitend op het vorige punt zal het organiseren van regelmatige audits zeker meerwaarde toevoegen. Audits zijn “onderzoek naar het functioneren van een bedrijf als geheel of op onderdelen.” (woorden.org, sd).

De genoemde verbeteringen zullen in de meeste gevallen een positieve impact maken op een bedrijf of organisatie. De verbeteringen zijn generiek en kunnen zowel in kleine als in grote schaal opgenomen worden. Het is vooral belangrijk om de implementatie goed uit te denken en te beginnen bij awareness van medewerkers, management en eventueel klanten.

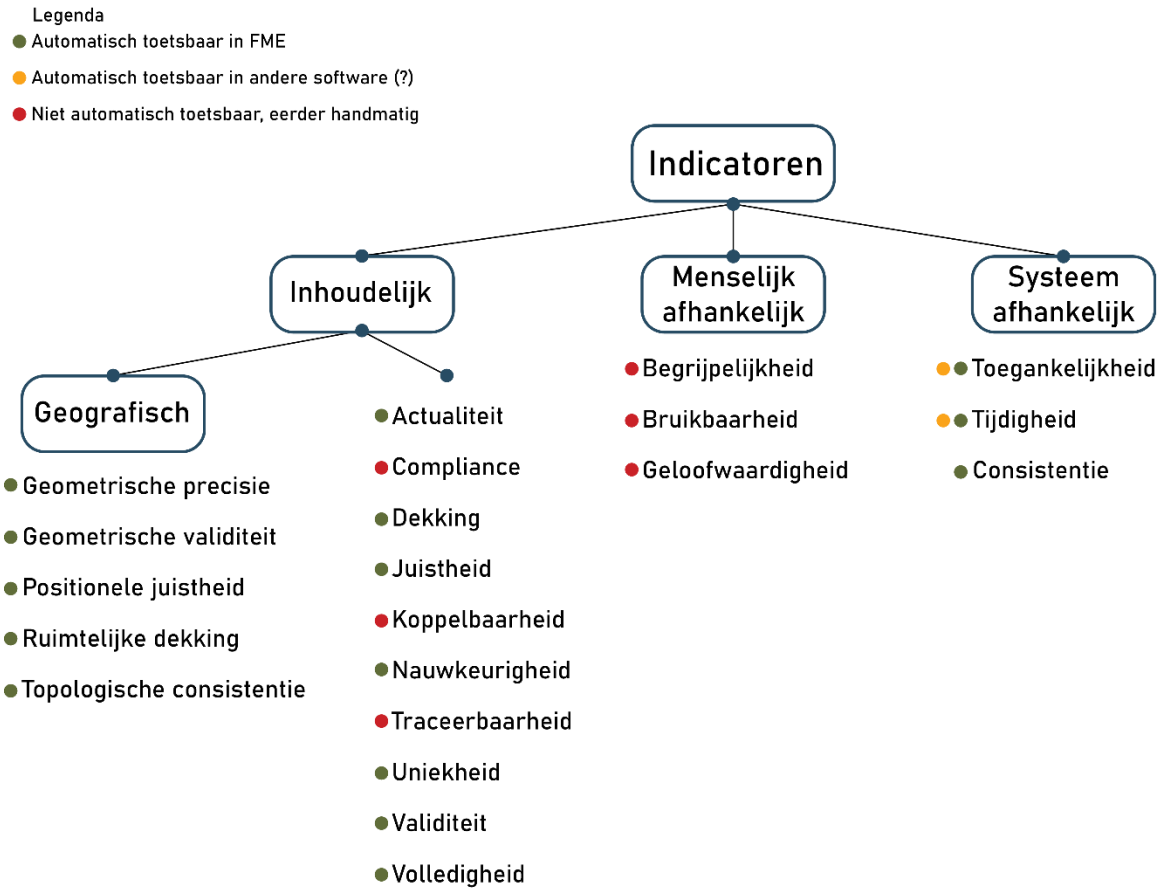
Zelfs als een bedrijf of organisatie alle stappen van standaardisatie, validatie en monitoring doorloopt is er nog geen garantie voor een optimale datakwaliteit. Alleen wanneer business-regels en indicatoren door audits constant getoetst en verbeterd worden biedt het meer zekerheid richting een optimale datakwaliteit.

6. Automatisch toetsbare indicatoren

De genoemde indicatoren in figuur 1 zijn hier overgenomen, maar dit keer met rode, gele en groene stippen. In de legenda staat kort uitgelegd wat deze stippen betekenen die voor de indicatoren staan.
Een groene stip geeft aan dat de desbetreffende indicator automatisch toetsbaar is in FME.
Een gele stip geeft aan dat de desbetreffende indicator waarschijnlijk automatisch toetsbaar is in een andere software. Hierbij wordt geen specifieke software genoemd omdat dat niet bekend is. Er is binnen dit onderzoek alleen gekeken naar de automatische toetsing in FME.
Een rode stip geeft aan dat de desbetreffende indicator niet automatisch toetsbaar is, maar juist beter eerder handmatig getoetst kan worden.

De indicatoren met de stoplicht stippen vormen een raamwerk dat ontwikkeld is voor dit onderzoek. Het doel van dit raamwerk is om meer variatie te creëren binnen de raamwerken en om een andere invalshoek te geven aan de toetsing van datakwaliteit en definities. Daarnaast kan het ook de fundering zijn van een Data Quality Control Platform.

Het raamwerk is slechts een verkenning en laat voorlopige resultaten zien. De aanduiding of het automatisch toetsbaar is of niet is op basis van de huidige kennis en inzicht in FME.



Figuur 2 Indicatoren van datakwaliteit met toelichting automatische toetsing

Verschil met eerdere raamwerken

De plaat van AAT

'De plaat van AAT' is een raamwerk gemaakt door Kasper Kisjes in 2016. Het is een raamwerk gemaakt voor Rijkswaterstaat om hun gegevens (automatisch) te kunnen toetsen. De plaat van AAT is het eerste raamwerk dat bij de start van dit onderzoek bekeken werd en waar veel inspiratie vandaan kwam.

Het verschil met dit andere raamwerk wordt benoemd om de waarde van de verschillende raamwerken in te zien en om meer diversiteit te bieden. Beide raamwerken kunnen gecombineerd worden om voor het bestemde doel een ideaal raamwerk te creëren.

De reden dat er binnen dit onderzoek een eigen raamwerk is gecreëerd is omdat er andere indicatoren gebruikt worden dan in het raamwerk van de plaat van AAT. Daarnaast wordt er in het raamwerk van de plaat van AAT gekeken naar automatische toetsing door middel van SAS dataflux. In het raamwerk van dit rapport wordt gekeken naar automatische toetsing door middel van FME. De laatste en een van de belangrijkste redenen om een nieuw raamwerk te creëren is omdat er in de plaat van AAT weinig aandacht specifiek naar geo-data gaat. Gunneman GIS & Geomatics richt zich specifiek op het gebruik van geo-data dus is het belangrijk dat geo-data goed uitgelicht wordt.

DAMA-DMBOK(2)

Een ander bekend raamwerk is het DAMA-DMBOK(2) raamwerk dat breed wordt gebruikt binnen overheidsinstanties . Dit raamwerk is ontwikkeld door een groot aantal dataprofessionals die hier meer dan 4 jaar aan gewerkt hebben.

Het voorziet vooral in de volgende zaken met betrekking tot het organiseren van Data Management binnen een organisatie:

- Standaard definities
- Guiding principles
- Good Practices
- Scope en afbakening per kennisgebied
- Gemeenschappelijke issues (tussen kennisgebieden)-
- Een uitgebreide literatuurlijst voor verdere studie

Kort gezegd: Het framework (raamwerk) heeft als doel consensus te creëren over het wat, wie en waarom van Data Management en de verschillende kennisgebieden (Data Kitchen, 2018).

Figuur 3 DAMA-DMBOK2 Guide Knowledge Area Wheel

In figuur 3 is het 'DAMA-DMBOK2 Guide Knowledge Area Wheel' te zien. Het wordt ook wel de eerste stap richting een goed datamanagement genoemd.

Het verschil tussen het DAMA-DMBOK2 raamwerk en het raamwerk van dit onderzoek is als volgt: Het DAMA-DMBOK2 raamwerk heeft de focus liggen op consensus creëren over het wat, wie en waarom van Data Management. Het raamwerk van dit onderzoek heeft de focus op (automatische) toetsing van indicatoren liggen.

Het DAMA-DMBOK2 raamwerk en het raamwerk van dit rapport kunnen daarentegen wel goed gecombineerd worden. Het DAMA-DMBOK2 raamwerk leent zich uitstekend voor het verbeteren van de data governance binnen het bedrijf of de organisatie. Het stukje 'Data Quality' van DAMA-DMBOK2 zoomt in op de datakwaliteit, iets waar het raamwerk van dit rapport bij kan ondersteunen door de toevoeging van (automatische) toetsing van indicatoren.

Er zijn enkele andere raamwerken, maar de plaat van AAT en het DAMA-DMBOK2 raamwerk zijn zeker noemenswaardig. Zoals genoemd kunnen raamwerken gecombineerd worden om voor een organisatie of bedrijf het ideale raamwerk te creëren. Stel de juiste prioriteiten op basis van de behoeften en doelen van het bedrijf of de organisatie.

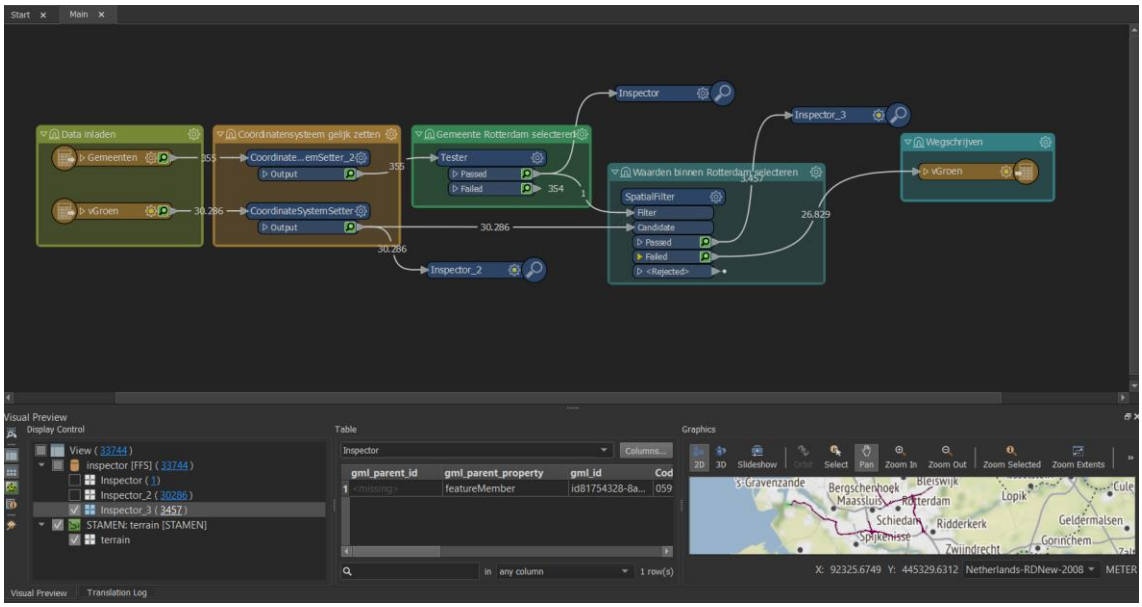
Automatisch toetsbare indicatoren in FME

Binnen dit onderzoek wordt een voorzet gegeven voor het automatisch toetsen van de datakwaliteitsindicatoren in FME. Het idee achter automatische toetsing is dat het toetsen van datakwaliteitsindicatoren zo efficiënt en generiek mogelijk wordt.

FME wordt gebruikt omdat het een tool is die zorgt voor een hoge reproduceerbaarheid. Net als bij een script kan FME opgeslagen worden en opnieuw geopend worden om later te gebruiken. Wanneer nieuwe data wordt toegevoegd aan de database wordt dit bij het opnieuw uitvoeren meegenomen zonder dat er aanpassingen nodig zijn. Binnen FME zijn instellingen om een 'workspace' regelmatig automatisch uit te voeren. Bij toetsing van datakwaliteit kan dat van pas komen door aan te geven dat de data om de X tijd getoetst moet worden.

In figuur 12 wordt een voorbeeld getoond van een FME workspace die alle datapunten binnen de gemeente Rotterdam selecteert. Bij automatische toetsing zou bijvoorbeeld getoetst kunnen worden op punten of vlakken buiten een bepaalde regio die daar niet horen.
De kleurenvensters om de 'transformers' (blauwe vakjes) geven aan wat er per onderdeel gebeurt. De eerste twee gele kadertjes zijn de inputwaarden. De ene inputwaarde zijn de gemeenten van Nederland en de andere inputwaarde is een dataset met diverse groenvlakken

van Nederland. De coördinatensystemen van beide inputwaarden worden gelijk gezet zodat daar geen afwijkingen voor kunnen komen. Bij de transformer ‘tester’ wordt gemeente Rotterdam uit alle gemeenten geselecteerd om verder mee te werken. De twee inputwaarden komen bij elkaar om alle waarden binnen Rotterdam te selecteren. Als laatst worden de waarden die niet binnen Rotterdam vallen weggeschreven naar een Excel sheet aangeduid door een foutmelding met toelichting.



Figuur 4 FME workspace om alle datapunten binnen gemeente Rotterdam te selecteren

Er zijn een hoop mogelijkheden met FME en er is al geëxperimenteerd met mogelijkheden voor automatische toetsing zoals hierboven uitgelegd aan de hand van figuur 12. Voor een aantal andere indicatoren is de automatische toetsing ook uitgewerkt. Binnen dit rapport wordt daar niet verder op ingegaan.

7. De kijk van experts op datakwaliteit, indicatoren van datakwaliteit en de automatische toetsing daarvan

Na het houden van zo'n achttien interviews ontstaat er een concreter beeld over de gedachten en meningen die experts/professionals hebben rondom het onderwerp datakwaliteit. Deels is al verwerkt in het rapport, maar er zijn nog een aantal noemenswaardige aspecten die uit de interviews naar voren kwamen. Vele uitspraken komen met elkaar overeen, maar er zijn een aantal genoemde aspecten die opvallen doordat ze nog niet eerder gehoord zijn binnen dit onderzoek, of omdat het een frisse blik werpt op het onderwerp.

In een gesprek met E. Franssen werd benoemd dat juistheid als indicatoren getoetst kan worden door middel van data lineage. Dat houdt in dat er een referentiecheck wordt gedaan door middel van een route van bron naar analyse af te leggen. Dit zorgt voor meer transparantie.

G. Bouwman heeft in een interview aangegeven dat wanneer de data niet in orde is, de data teruggestuurd moet worden in plaats van opgepoetst. Data leveranciers kunnen vrij laks worden en ervan uitgaan dat de mensen die de data verwerken ook zullen opschonen.

In een gesprek met P van Wee werd genoemd dat schonen van data alleen nuttig is als er iemand is die het belangrijk vindt. Als dat niet de eigenaar van de data is maar bijvoorbeeld de ontvanger dan is het van belang dat de twee partijen tot de afspraak komen dat, en hoe er geschoond gaat worden. Dat schonen gebeurt veelal in opdracht/aansturing van de data eigenaar. Ga pas aan de slag als er consensus is over opschoning bij alle betrokken partijen.

In het gesprek met W. Tadema kwam naar voren dat de service level agreements (SLA) van een machine learning systeem vaak alleen gebaseerd zijn op beschikbaarheid en performance (bijvoorbeeld de snelheid waarmee het systeem een voorspelling retourneert). Dat is niet voldoende, de kwaliteit van de voorspellingen moet namelijk ook gemonitord worden. Er kan veel tijd zitten tussen het doen van een voorspelling en de daadwerkelijke uitkomst (het moment waarop geverifieert wordt of de voorspelling juist was). Soms is het helemaal niet mogelijk om de uitkomst te verifiëren, omdat je door de wijze waarop je ingrijpt op het proces niet meer kunt achterhalen wat er gebeurt zou zijn als je dat niet had gedaan.

Om deze reden is het heel belangrijk om ook de input data van het machine learning model te monitoren, en wel op twee manieren.

Ten eerste moet de datakwaliteit van de input van het machine learning model continu gemonitord worden. Voor machine learning modellen geldt 'garbage in, garbage out'. Als de datakwaliteit van de input slecht is, kan de output van het model nooit goed zijn.

Daarnaast is het noodzakelijk om veranderingen in de input data te detecteren. Het gaat in dit geval dus niet om fouten of andere onvolkomenheden in de data, maar bijvoorbeeld om een verschuiving in de verdeling. Het machine learning model is namelijk getraind op een dataset met bepaalde karakteristieken. Het model kan alleen betrouwbare voorspellingen geven zo lang de data waarop het model getraind is, representatief is voor de populatie waarop het model wordt toegepast. Als er een verschuiving is binnen de populatie, moet dat tijdig gedetecteerd worden, zodat het model opnieuw getraind kan worden op data die wel representatief is voor de nieuwe situatie. Ter illustratie: vermoedelijk zullen er door de coronamaatregelen heel wat modellen zijn die nu niet meer goed voorspellen, omdat ze gebaseerd zijn op een werkelijkheid die niet meer bestaat.

Wanneer een model slechter gaat presteren, moet dus gekeken worden naar de input data: is de datakwaliteit afgenomen of is de trainingsdata niet meer representatief?

Daarnaast benoemde W. Tadema de indicator ‘koppelbaarheid’. De indicator koppelbaarheid zorgt voor het gemakkelijk joinen (combineren) van data. Voeg bijvoorbeeld een bag ID toe aan de dataset voor het makkelijk joinen van data. Deze indicator is toegevoegd aan het raamwerk.

A. Klink geeft in een gesprek aan dat om betrouwbaarheid te onderbouwen er een rekensommetje gemaakt zou kunnen worden van alle afwijkingen die tijdens het verwerken en inwinnen zijn opgetreden. Dit geeft een beter beeld van de betrouwbaarheid van de data en zorgt voor minder fouten bij doorberekeningen.

8. Enquêteresultaten onderzoek datakwaliteit

Na het afnemen van de enquête zijn de antwoorden van de respondenten geanalyseerd. Uit de uitkomsten van deze analyses zijn grafieken en diagrammen gemaakt ter ondersteuning van de getrokken conclusies.

Alle respondenten die er minder dan 2 minuten over hebben gedaan om de enquête in te vullen zijn gecontroleerd op antwoorden. De reden hierachter is omdat 2 minuten vrij kort is voor het invullen van de enquête en er zekerheid moet zijn dat de antwoorden niet 'nep' of expres vervalst zijn. Uiteindelijk zijn er geen respondenten verwijderd omdat er geen abnormale antwoorden zijn gegeven. Respondenten kunnen de enquête ook maar één keer invullen.

In totaal zijn er 366 respondenten die de enquête hebben ingevuld. Bij elke vraag staat 'N = x' om het aantal respondenten aan te geven dat die vraag heeft ingevuld. Het kan zo zijn dat er minder respondenten bij een vraag staan. Dit komt door 'question branching', ofwel vertakkingen van vragen. Sommige antwoorden op vragen zorgen ervoor dat de respondent een vraag overslaat of juist een extra vraag in moet vullen.

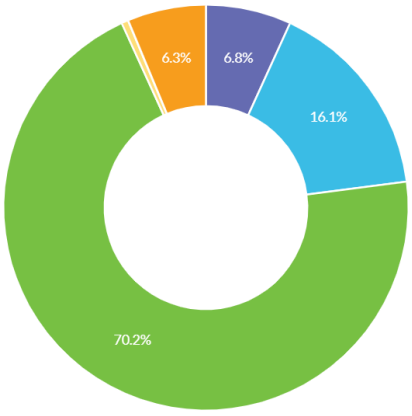
Let op dat er bij enkele vragen meerdere antwoorden gekozen konden worden door de respondenten. De optelsom van de antwoorden zal daarom niet altijd op n = x uitkomen. Let ook op dat bij de toelichting niet altijd alle uitkomsten uitgelicht worden. Om het rapport efficiënt leesbaar te houden worden alleen de relevantste uitkomsten genoemd. Daarnaast wordt er vervolgonderzoek uitgevoerd om verbanden tussen vragen uit te lichten en meer conclusies te trekken, deze zijn nu nog niet uitgebreid.

Algemene vragen

Door de uitkomsten van de algemene vragen kan er een profiel gebouwd worden voor de respondenten. Hierdoor wordt onze bereikte groep respondenten duidelijk en wordt er duidelijk of er genoeg variatie in de uitkomsten zit. Daarnaast zijn de algemene vragen ook nuttig om verbanden tussen antwoorden en gebruikers te vinden.

Q1. Met wat voor data werk je? (N = 366)

Een ruime 70% van de respondenten geeft aan zowel met geodata als met 'gewone' data te werken. Zo'n 16% geeft aan dat zij alleen met geodata werken en een ruime 6% viel af doordat zij niet met data werken.



Choice	Total
met "gewone" data. (De data bevat geen geografische coördinaten)	25
met geodata. (De data bevat geografische coördinaten)	59
zowel met "gewone" data als met geodata. (De data kan geografische coördinaten bevatten)	257
met data maar ik weet niet of dat dat "gewone" data of geodata is.	2
ik werk niet met data	23

Figuur 5 Resultaten in een grafiek voor Q1

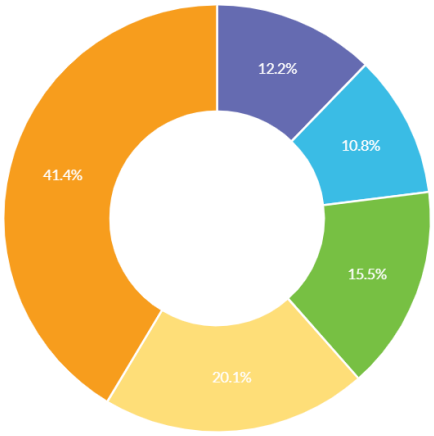
Q2. Wat zijn je taken met data? (N = 343)

Bij deze vraag was het mogelijk om meerdere antwoorden aan te kruisen. Het aantal keren dat het antwoord is aangekruist is gedeeld door het totaal aantal respondenten van de vraag, maal 100. Het percentage wat hieronder weergegeven wordt is het totale percentage van alle respondenten dat de vraag heeft aangekruist.

Taken	Percentage (hoog naar laag)
visualiseren van data	71,43%
analyseren van data voor rapportage	64,72%
zorgen voor goed gebruik van data in organisaties	53,35%
automatiseren van verwerking van data	52,77%
controleren van data/fouten opsporen	49,56%
beheren van data	49,27%
inwinnen van data	32,94%
(in)tekenen van data (denk aan AutoCAD/GIS)	30,61%
invoeren van data	30,03%
analyseren van data voor wiskundige berekeningen/machine learning	27,70%
anders, namelijk	9,91%

Q3. Hoe beoordeel je jezelf als het gaat om jouw expertise met betrekking tot data? (N = 343)

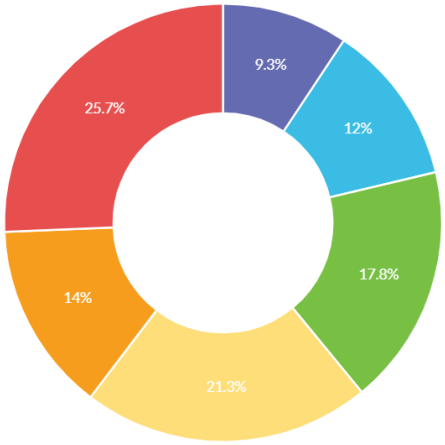
40% van de respondenten geeft aan zichzelf te beoordelen als senior als het gaat om de expertise met betrekking tot data. Dit is bijna twee keer zo veel als de respondenten die aangaven tussen medior en senior in te zitten, namelijk zo'n 20%. Toch is de variatie binnen de groep respondenten hier vrij goed verdeeld op de vele seniors na.



Choice	Total
junior	42
tussen junior / medior	37
medior	53
tussen medior / senior	69
senior	142

Figuur 6 Resultaten in een grafiek voor Q3

Q4. Hoeveel uren per week ben je (ongeveer) bezig met data? (N = 343)
Ruim een kwart van de respondenten geeft aan meer dan 32 uur per week met data bezig te zijn. Bijna 10% geeft aan dat zij tussen de 0 en 4 uur bezig zijn met data.



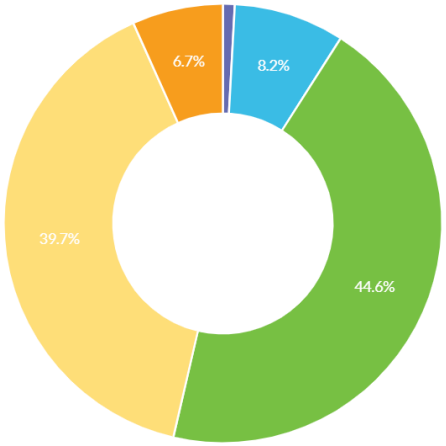
Choice	Total
tussen de 0 en 4 uur	32
tussen de 4 en 8 uur	41
tussen de 8 en 16 uur	61
tussen de 16 en 24 uur	73
tussen de 24 en 32 uur	48
meer dan 32 uur	88

Figuur 7 Resultaten in een grafiek voor Q4

Vragen datakwaliteit

Q5. Welke mate van inzicht in de kwaliteit van data denk je te hebben binnen je huidige werkgebied? (N = 343)

Minder dan 1% van de respondenten geeft aan geen inzicht te hebben in de kwaliteit van data. Bijna 45% geeft aan redelijk inzicht te hebben in de kwaliteit van data. Bijna 40% geeft aan veel inzicht te hebben in de kwaliteit van data en zo'n 7% geeft aan volledig inzicht te hebben in de kwaliteit van data. Wat hieruit geconcludeerd kan worden is dat de algemene kennis van de data onder de respondenten vrij hoog is.
Deze conclusie kan echter niet te snel getrokken worden, het gaat namelijk over de mate van inzicht die respondenten denken te hebben. Iemand met veel verstand van datakwaliteit kan inzien dat er weinig inzicht in de datakwaliteit is, terwijl iemand die weinig verstand heeft van datakwaliteit denkt dat er een hoop inzicht is. Het ligt er maar net aan waar de lat ligt.



Choice	Total
geen inzicht in de kwaliteit van data	3
weinig inzicht in de kwaliteit van data	28
redelijk inzicht in de kwaliteit van data	153
veel inzicht in de kwaliteit van data	136
volledig inzicht in de kwaliteit van data	23

Figuur 8 Resultaten in een grafiek voor Q5

Q6. Vind je het belangrijk om inzicht in de kwaliteit van data te hebben? (N = 343)

De conclusie trekken uit deze vraag is vrij gemakkelijk. 99% van de respondenten geeft aan dat ze het belangrijk vinden om inzicht te hebben in de kwaliteit van data. De 1% die 'nee' heeft geantwoord **(Q7)** (N = 4) geeft aan nog geen urgente behoefte gehad te hebben, geen ruimte te hebben om te investeren in betere kwaliteit van data of te weinig met data te werken.

Q8. Ondervind je problemen in je werk doordat de kwaliteit van data onvoldoende is? (N = 343)

Ruim driekwart van de respondenten geeft aan problemen te ondervinden doordat de kwaliteit van data onvoldoende is. Dit is een heel hoog percentage van de mensen en geeft dus aan dat onvoldoende datakwaliteit in veel gevallen zorgt voor problemen.

Q9. Welke problemen komen voor binnen de kwaliteit van data? (N = 258)

Verouderde data en missende metadata lijken een van de grootste probleem veroorzakers binnen de datakwaliteit. Met de ID's lijkt minder vaak wat mis, maar alsnog hebben ongeveer 20% van de respondenten daar last van.

Problemen binnen kwaliteit van data	Percentage (hoog naar laag)
de data is verouderd	65,89%
de data mist metadata	63,57%
het is onduidelijk waar gegevens vandaan komen	50,78%
de data bevat lege velden	49,22%
relaties met andere gegevens kloppen niet	45,74%
geometrie van getekende objecten klopt niet	36,82%
locaties kloppen niet	34,88%
topologische problemen zoals overlap of gaten komen voor	34,11%
gegevens voldoen niet aan (inter)nationale standaarden zoals de basisregistraties	28,68%
de data bevat te veel informatie waardoor het veel tijd aan zoekwerk kost	25,58%
domeinwaarden kloppen niet	21,32%
ID's missen waardoor gegevens lastig te traceren zijn	20,93%
ID's komen dubbel voor	18,99%
anders, namelijk	13,95%

Q10. Noem maximaal 3 oorzaken waardoor de datakwaliteit onvoldoende is. Je hoeft niets in te vullen als je het niet weet. (N = 204)

In de ruwe data kunnen alle volledige antwoorden gevonden worden.

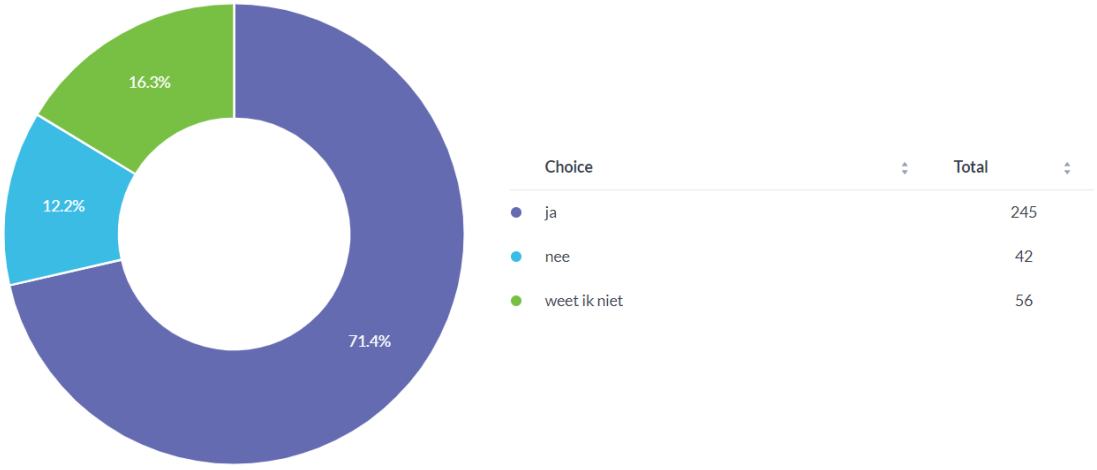
Q11. Hoe bewust is men binnen jouw organisatie van datakwaliteit en de eventuele gevolgen van een onvoldoende kwaliteit van data? (N = 343)

Links (1): geheel niet bewust van datakwaliteit – Rechts (7): volledig bewust van datakwaliteit

De bewustheid van de organisaties van respondenten ligt vrij hoog. Zo is het gemiddelde 5 op en schaal van 1 tm 7 met de meeste antwoorden op een 6.
Dit is positief, zoals in het onderzoek naar voren komt is bewustheid stap nummer één richting een goede datakwaliteit.

Q12. Worden er stappen gezet om het bewustzijn rondom het belang van voldoende datakwaliteit te vergroten? (N = 343)

Een ruime 70% geeft aan dat zij stappen zetten om het bewustzijn rondom het belang van voldoende datakwaliteit te vergroten. Ongeveer 12% geeft aan dat niet te doen en de overige respondenten geven aan dat niet te weten.



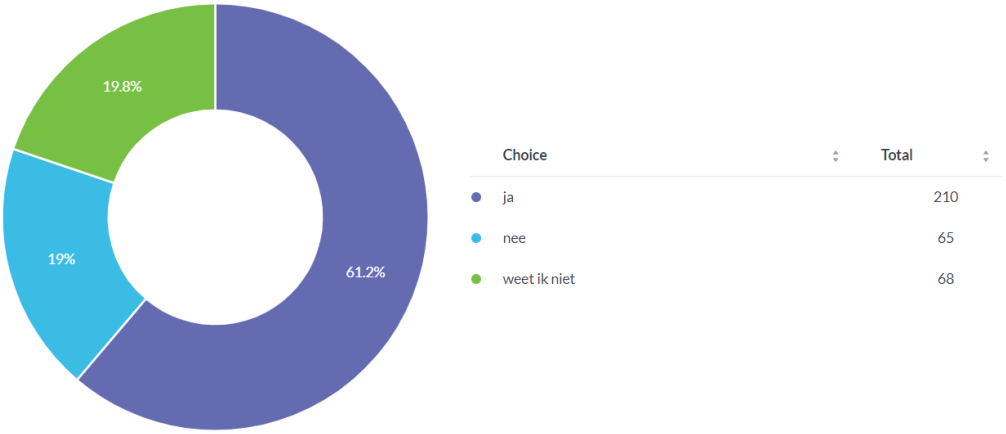
Figuur 9 Resultaten in een grafiek voor Q12

Q13. Wat voor stappen zijn dat? (N = 244)

In de ruwe data kunnen alle volledige antwoorden gevonden worden.

Q14: Wordt de kwaliteit van gegevens binnen je organisatie getoetst? (Door bijvoorbeeld fouten te melden of te rapporteren in de metadata) (N = 343)

Een ruime 60% van de respondenten toetst al gegevens binnen de organisatie voor de kwaliteit. Zo'n 20% toetst niet op kwaliteit van gegevens en de laatste 20% weet het niet.



Figuur 10 Resultaten in een grafiek voor Q14

Stappen zetten rondom het bewustzijn van het belang van voldoende datakwaliteit ligt zo'n 10% hoger. Een mogelijke oorzaak is dat stappen zetten rond bewustzijn eerder komt dan het daadwerkelijk toetsen van de gegevens en dat de organisatie daar nog niet aan begonnen is. Een andere mogelijke oorzaak is dat de organisatie de middelen niet heeft om te toetsen.

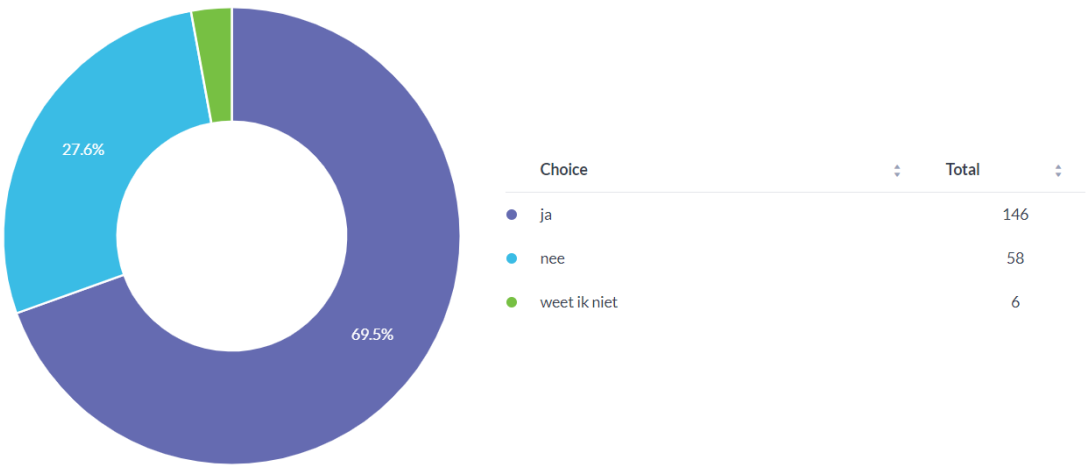
Q15. Hoe wordt er inzicht verkregen in de datakwaliteit? (N = 210)

De manier waarop vooral inzicht wordt verkregen is door middel van het intern melden van fouten en daarna door onduidelijkheden over de data(kwaliteit) te overleggen. Dit zijn vrij laagdrempelige manieren om inzicht te krijgen, maar helpen wel om al awareness te creëren binnen de organisatie. Deze manier van inzicht vereist ook niet direct een systeem en kan simpelweg uitgesproken worden. De andere manieren van inzicht die digitaal gaan worden beduidend minder gekozen.

Manieren van inzicht in datakwaliteit	**Percentage (hoog naar laag)**
Fouten worden intern gemeld	73,81%
Bij onduidelijkheden over de data(kwaliteit) wordt overleg gevoerd	67,14%
Afwijkingen of fouten worden geregistreerd (bijvoorbeeld in excel of als extra kolom in een tabel)	42,86%
Fouten worden extern gemeld	39,52%
Informatie over de datakwaliteit wordt opgenomen in de metadata	29,52%
Trendanalyses worden uitgevoerd	28,57%
Anders, namelijk	15,24%
Er wordt geen inzicht verkregen in de kwaliteit van data	2,86%

Q16. Speel je zelf een actieve rol in de toetsing van datakwaliteit binnen je organisatie? (N = 210)

Bijna 70% van de respondenten geeft aan zelf een actieve rol te spelen in de toetsing van datakwaliteit binnen hun organisatie. Ruim een kwart van de respondenten geeft aan dat niet te doen en een kleine 3% weet het niet.



Figuur 11 Resultaten in een grafiek voor Q16

Q17. Een aantal aspecten van datakwaliteit kunnen automatisch getoetst worden. Gebeurt dat in jouw organisatie? (N = 210)

Een kleine 70% van de respondenten geeft aan dat een aantal aspecten van datakwaliteit automatisch getoetst worden binnen hun organisatie. Een kleine 20% geeft aan dat dat niet gebeurt en een ruime 10% geeft aan dat niet te weten.

Q18. Welke aspecten worden (automatisch) getoetst? (N = 145)

In de ruwe data kunnen alle volledige antwoorden gevonden worden.

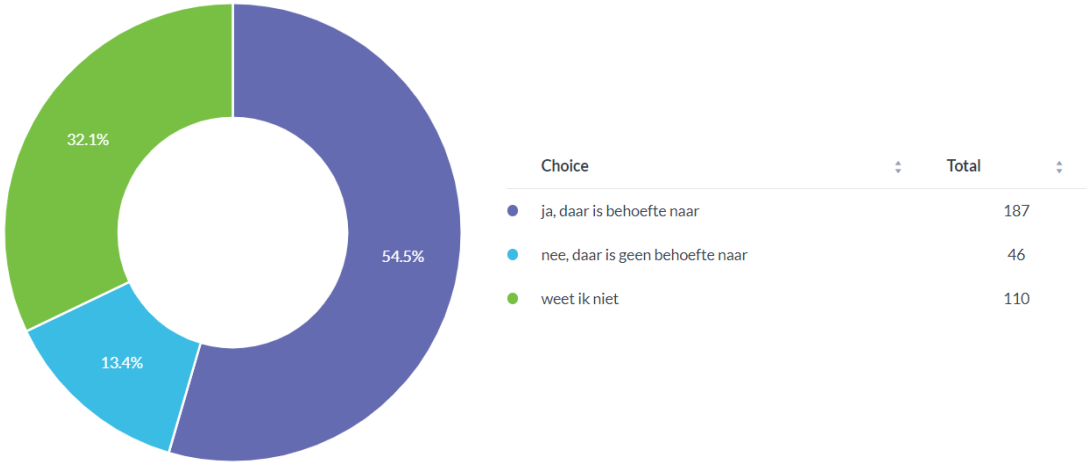
Q19. Is er behoefte naar (automatische) toetsing van aspecten? (N = 343)

Zo'n 55% van de respondenten geeft aan dat er behoefte is naar (automatische) toetsing van aspecten. Zo'n 13% geeft aan dat er geen behoefte naar (automatische) toetsing van aspecten. Als laatst geeft zo'n 32% van de respondenten aan niet te weten of er behoefte is naar (automatische) toetsing van aspecten.

Een verklaring voor de grote hoeveelheid respondenten die 'weet ik niet' hebben geantwoord kan zijn dat het in hun ogen nog te abstract is. Vooral automatische toetsing is iets waar weinig mensen een beeld bij kunnen vormen, dat bleek ook uit de interviews.

Vormen van inzicht in de datakwaliteit	**Percentage (hoog naar laag)**
door gebruik van een Viewer, GIS, of Dashboard om geografische data (met afwijkingen of fouten) te bekijken	61,62%
door gebruik van een meldingen-Dashboard	45,45%
door gebruik van een statistieken-Dashboard om bijvoorbeeld trends te kunnen zien	44,78%

door per datakwaliteitsindicator te zien wat de status van de gegevens zijn	44,44%
door per KPI te zien wat de status van de gegevens zijn	29,29%
door meldingen van afwijkingen of fouten in een (excel) spreadsheet	28,96%
door gebruik van een API om afwijkingen en fouten te downloaden	27,61%
anders, namelijk	9,76%



Figuur 12 Resultaten in een grafiek voor Q9

Q20. In welke vorm zou je het liefst inzicht krijgen in kwaliteit van data? (N = 297)

De top drie keuzes voor de vorm van inzicht in de kwaliteit van data zijn allemaal dashboard gericht.

Q21. Welke datakwaliteitsindicatoren of KPI's zijn voor jou het belangrijkst om inzicht in te krijgen? (N = 296)

In de ruwe data kunnen alle volledige antwoorden gevonden worden.

Q22. Heb je nog opmerkingen over deze enquête? (N = 341)

In de ruwe data kunnen alle volledige antwoorden gevonden worden.

Discussie

Binnen het onderzoek is de validiteit op meerdere manieren aangetoond. Allereerst zijn de enquêteresultaten als open data beschikbaar, dit maakt het onderzoek reproduceerbaar. Daarnaast zijn vrijwel alle gebruikte bronnen recent en komen ze van overheidsinstanties, professionals in het vakgebied en websites die aan correcte bronvermelding voldoen. Als laatst zijn de interviews afgenomen onder een diverse doelgroep van verschillende beroepen rondom data.

De resultaten van het onderzoek komen redelijk overeen met de gestelde verwachtingen vóór het onderzoek. Voorafgaand is er middels literatuuronderzoek al enig inzicht verkregen. Een specifiek resultaat dat niet aan de gestelde verwachtingen deed kwam uit de enquête. Het aantal respondenten dat aangaf dat er getoetst wordt op datakwaliteit viel hoger uit dan de verwachtingen. Uit literatuuronderzoek kwam vaak naar voren dat toetsing van datakwaliteit binnen bedrijven weinig voorkwam. Uit de resultaten van de enquête bleek anders.

Er zijn een hoop nieuwe inzichten verkregen door het afnemen van de interviews en gesprekken met professionals. Deze inzichten geven een duidelijk beeld over hoe er wordt omgegaan met datakwaliteit binnen organisaties en bedrijven.

Met de resultaten van dit onderzoek is aangetoond dat automatische toetsing van indicatoren mogelijk is, maar ook dat er behoefte naar is. Daarnaast toont dit onderzoek aan dat er veel belang is voor datakwaliteit binnen organisaties, maar dat vervolgstappen soms ingewikkeld kunnen zijn.

Een mogelijke kanttekening bij het onderzoek is de betrekkelijk korte tijd waarin het onderzoek geschreven is. Door een korte tijdsspan kan het zijn dat er informatie mist die wel degelijk van belang is.

Het onderzoek is slechts een aanzet tot verder onderzoek en de ontwikkeling van een ‘Data Quality Control Dashboard’. Een vervolgonderzoek voegt meerwaarde toe, er zijn namelijk een aantal aspecten die uitgebreider behandeld kunnen worden. Zo missen er Chi^2-testen om verbanden statistisch te verklaren tussen antwoorden van respondenten en zouden er uitgebreidere conclusies gegeven kunnen worden. In vervolgonderzoek zou dieper ingegaan kunnen worden op het combineren van diverse raamwerken om een beter totaalplaatje te krijgen en zou er verdieping in FME of andere software gemaakt kunnen worden voor een meer technische kant van het onderzoek. Als laatste zou in vervolgonderzoek ook dieper ingegaan kunnen worden op kwaliteit van data governance, de kwaliteit van processen, de kwaliteit van systemen en de kwaliteit van informatie modellen aangezien deze een belangrijke rol spelen binnen het kader van datakwaliteit.

Conclusie

Om een conclusie te kunnen trekken wordt de hoofdvraag nog een keer beschreven:
Hoe kunnen indicatoren van datakwaliteit vertaald worden naar (automatische) toetsing van datakwaliteit?

Het is lastig om te zeggen of er één juiste definitie van datakwaliteit is. Wat belangrijk is, is dat er gekeken wordt naar het desbetreffende bedrijf of de organisatie en de data waar mee gewerkt wordt. Er moet een onderbouwde keuze gemaakt worden in de definitie voor datakwaliteit die aansluit bij de wensen van een organisatie of bedrijf.

Iedereen heeft direct of indirect baat bij een goede datakwaliteit. Het is belangrijk dat zoveel mogelijk mensen die betrekking hebben tot de data de meerwaarde inzien van een verbeterde datakwaliteit.

De belangrijkste oorzaken van een slechte datakwaliteit zijn; onnauwkeurigheid bij de notitie van metadata en bronnen, afwijking in meetapparatuur of verkeerd gebruik van meetapparatuur, onvoldoende standaardisatie, verkeerde boorberekeningen van attributen, onvoldoende gevoel tussen de relatie van gegevens en de werkelijkheid en data die ten onrechte wel of niet aan het systeem wordt toegevoegd. Daarnaast speelt de kwaliteit van data governance, de kwaliteit van processen, de kwaliteit van systemen en de kwaliteit van informatie modellen ook een belangrijke rol.
De positieve gevolgen van een betere datakwaliteit zijn; snellere en betere besluitvorming, efficiëntere en voorspelbaardere processen, compliance en een eventueel concurrentievoordeel. Dit alles zorgt voor meer tevreden stakeholders en minder triviale fouten.
De negatieve gevolgen van slechte datakwaliteit zijn; grotere risico's door het nemen van beslissing op basis van foutieve data, een lager werkniveau onder werknemers waar eventuele frustratie bij komt kijken en een slechter bedrijfsimago.
Als laatste kan te hoge datakwaliteit kan voor onnodig hoge kosten zorgen.

Indicatoren moeten gestandaardiseerd zijn binnen een organisatie om verwarring te voorkomen. Datakwaliteitsindicatoren maken datakwaliteit concreter. Welke indicatoren relevant zijn voor een bedrijf of organisatie ligt aan de waarde die gehecht wordt aan een indicatoren en of de kosten opwegen tegen de baten.
Om indicatoren meetbaar te maken kunnen KPI's aan indicatoren gehangen worden met absolute of relatieve waarden. Een andere manier is door middel van outlier modellen.

Het verbeteren van de datakwaliteit kan op een aantal manieren; maak helder dat datakwaliteit een constant lopend project is, maak helder wie het eigenaarschap van de datakwaliteit draagt, leg alleen relevante data vast, kies de juiste indicatoren en stel prioriteiten, dwing invoer van data technisch af, gebruik eventueel raamwerken, spreek standaarden af voor externe data, zorg voor transparantie binnen de organisatie rond de data, monitor de data met een nulmeting en stel audits op.

Raamwerken kunnen gecombineerd worden om voor een organisatie of bedrijf het ideale raamwerk te creëren. Stel de juiste prioriteiten op basis van de behoeften en doelen van het bedrijf of de organisatie.

De enquêteresultaten laten concreet zien dat er grote belangstelling is voor datakwaliteit en de toetsing ervan. Wat ook naar voren komt is dat een hoop respondenten al enige vorm van toetsing binnen hun organisatie uitvoeren, met name het intern melden van fouten.

Om te concluderen en antwoord te geven op de hoofdvraag;
Definities van datakwaliteitsindicatoren moeten binnen bedrijven gestandaardiseerd worden door middel van een raamwerk of andere vorm van schema. Niet alle datakwaliteitsindicatoren en aspecten hoeven meegenomen te worden. Belangrijk is om inzicht te verkrijgen in het doel waar de data voor dient en welke behoeftes er zijn en op basis daarvan aan de slag te gaan met (automatische) toetsing.

Verwijzingen

Black, A., & van Nederpelt, P. (2020). *Dictionary of dimensions of data quality (3DQ)*. Opgehaald van DAMA NL: http://www.dama-nl.org/wp-content/uploads/2020/11/3DQ-Dictionary-of-Dimensions-of-Data-Quality-version-1.2-d.d.-14-Nov-2020.pdf

Bouman, E. (2005, Maart 23). *Testen van datakwaliteit* . Opgehaald van https://www.testnet.org/testnet/download/presentaties-thema-avond-datakwaliteit-23-maart-2005/testen_van_datakwaliteit.pdf

Conijn, F., De Jong, J., & Krol, M. (2019, december 30). *Datamanagement: kwaliteit gegevens bepaalt kwaliteit informatie*. Opgehaald van controllers magazine: https://cmweb.nl/2019/12/datamanagement-kwaliteit-gegevens-bepaalt-kwaliteit-informatie/?vakmedianet-approve-cookies=1&_ga=2.29341881.441421551.1599549760-618428468.1599549760

Data Kitchen. (2018, Maart 30). *Een gedeeld begrip van het DAMA-DMBOK raamwerk is de eerste stap in effectief Data Management.* Opgehaald van Data Kitchen: https://datakitchen.nl/data-nieuws/een-gedeeld-begrip-van-het-dama-dmbok-raamwerk-is-de-eerste-stap-in-effectief-data-management/

Ensie. (2011, Maart 1). *Kwaliteit*. Opgehaald van Ensie: https://www.ensie.nl/redactie-ensie/kwaliteit

Gartner. (2021). *Data Governance*. Opgehaald van Gartner: https://www.gartner.com/en/information-technology/glossary/data-governance

Gunneman, J. (2020). Uitspraak. (S. Mans, Interviewer)

Harmelink, R. (2016, December 8). *Datakwaliteit: hoe belangrijk is dat eigenlijk?* Opgehaald van Executive Finance: 2016

Heertum, H. v. (2020, Oktober 2). Interview datakwaliteit. (S. Mans, Interviewer)

Infomedics. (2017, juni 6). *Informatiemodel – wat is het en waarom nodig?* Opgehaald van Infomedics: https://www.infomedic.nl/informatiemodel/

Jutte, B. (2020, Oktober 5). Interview datakwaliteit . (S. Mans, Interviewer)

Kaper, J. (2020, September 9). Gesprek over datakwaliteit. (S. Mans, Interviewer)

Klaassen, N. (2004, November). *Datakwaliteit is meetbaar*. Opgehaald van biplatform.nl: https://biplatform.nl/magazines/Aveq/108433.pdf

Klink, A. (2020, September 9). Gesprek datakwaliteit . (S. Mans, Interviewer)

Legerstee, M. (2019, augustus 30). *Datakwaliteit: tien basistips voor het verhogen en waarborgen van je datakwaliteit*. Opgehaald van cmotions: https://cmotions.nl/datakwaliteit-tien-basistips/

Omnicsci. (sd). *Geodata*. Opgehaald van Omnisci: https://www.omnisci.com/technical-glossary/geodata

Redman, T. C. (2016, September 22). *Bad Data Costs the U.S. $3 Trillion Per Year*. Opgehaald van Harvard Business Review: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

Sogeti. (sd). *In 3 stappen naar goede datakwaliteit*. Opgehaald van Sogeti: https://www.sogeti.nl/nieuws/data-business-analytics/blogs/3-stappen-naar-goede-datakwaliteit

Voorter, R. (2020, Oktober). (S. Mans, Interviewer)

Weber, D. (2020, Augustus 27). *Het belang van goede datakwaliteit*. Opgehaald van Exsell: https://www.exsell.nl/het-belang-van-goed-datakwaliteit/

woorden.org. (sd). *audit*. Opgehaald van woorden.org: https://www.woorden.org/woord/audit