Onderzoek datakwaliteit

Resultaten enquête

Spijker Mans | Gunneman GIS & Geomatics | 02/02/2021

Inhoudsopgave

Inleiding

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 enquête is uitgezet om het kwalitatieve onderzoek naar datakwaliteit verder te verdiepen met concrete cijfers. Het onderzoek is uitgevoerd voor Gunneman GIS & Geomatics met als doel inzicht te krijgen in de marktwaarde van een datakwaliteit dashboard. Binnen dit rapport worden de resultaten en conclusies van de enquête over datakwaliteit besproken en toegelicht.

De enquête is opgezet door middel van 'SurveyPlanet', een online enquêtetool. Het uitzetten van de enquête is gedaan door middel van LinkedIn en persoonlijke benadering (zowel online als offline). Vermoedelijk valse respondenten zijn verwijderd en respondenten zijn maar eenmaal toegelaten om de enquête in te vullen. De eenmalige toelating gaat per device. Het aantal respondenten na het sluiten van de enquête en het verwijderen van eventuele valse respondenten eindigde op 366 (N=366).

Resultaten

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 per device.

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.

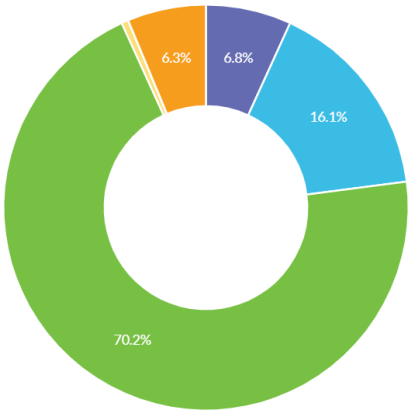
Let op dat er bij enkele vragen meerdere antwoorden gekozen konden worden door de respondenten. De optelsom van de antwoorden zal daarom niet altijd op n = x uitkomen. Let ook op dat bij de toelichting niet altijd alle uitkomsten uitgelicht worden. Om het rapport efficiënt leesbaar te houden worden alleen 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 1 RESULTATEN IN EEN GRAFIEK VOOR Q1

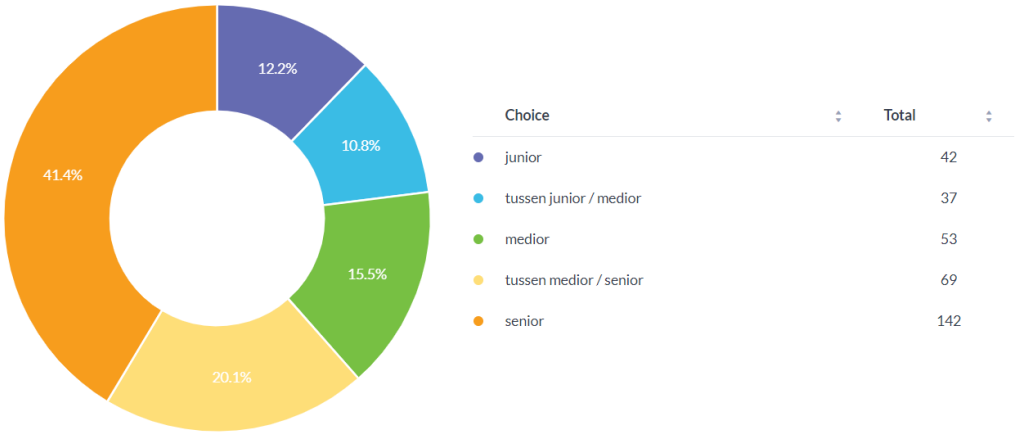
Q2. Wat zijn je taken met data? (N = 343)

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%

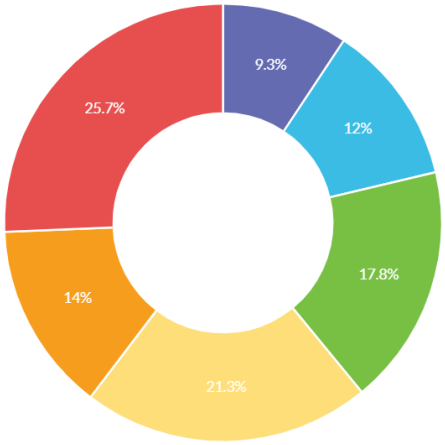
Q3. Hoe beoordeel je jezelf als het gaat om jouw expertise met betrekking tot data? (N = 343)

Ruim 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.



FIGUUR 2 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.



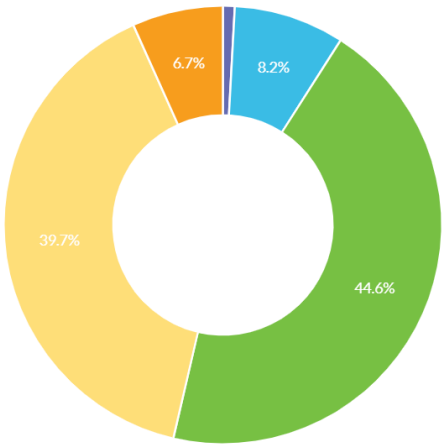
Choice	Total
tussen de 0 en 4 uur	32
tussen de 4 en 8 uur	41
tussen de 8 en 16 uur	61
tussen de 16 en 24 uur	73
tussen de 24 en 32 uur	48
meer dan 32 uur	88

FIGUUR 3 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 4 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)

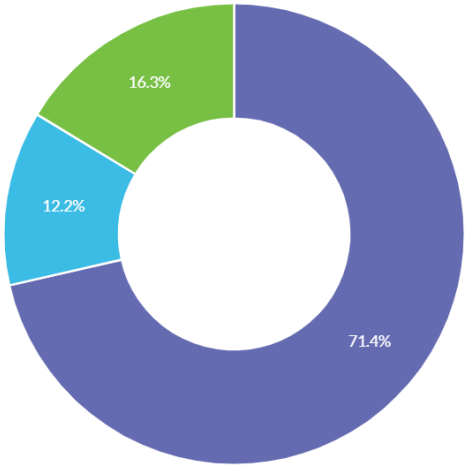
Links (1): geheel niet bewust van datakwaliteit – Rechts (7): volledig bewust van datakwaliteit

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.



Choice	Total
ja	245
nee	42
weet ik niet	56

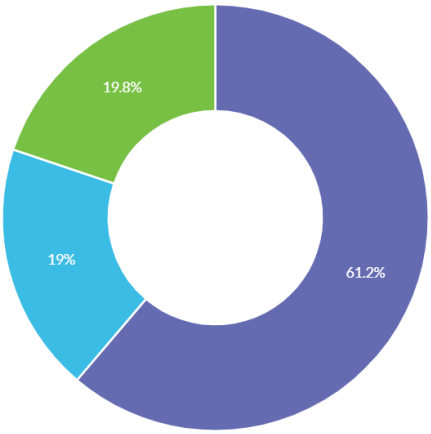
FIGUUR 5 RESULTATEN IN EEN GRAFIEK VOOR Q12

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.



Choice	Total
ja	210
nee	65
weet ik niet	68

FIGUUR 6 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.

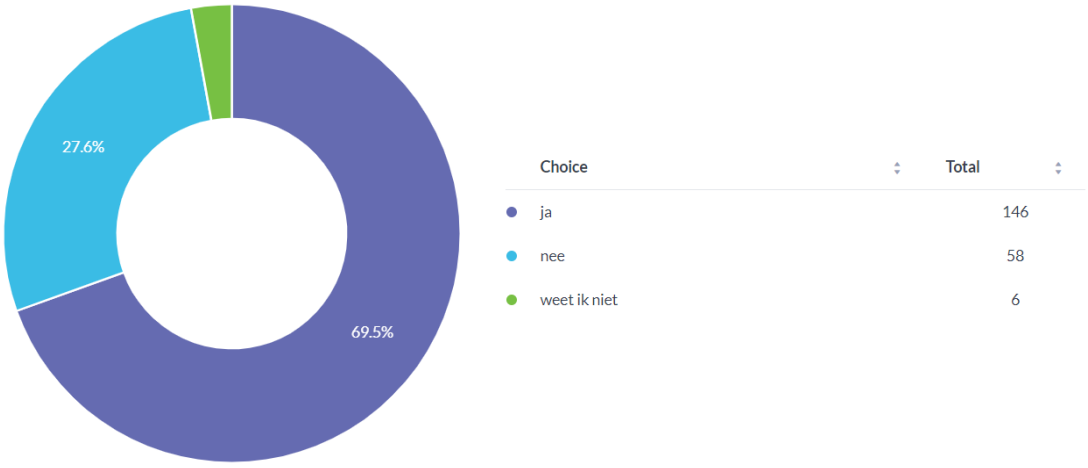
Q15. Hoe wordt er inzicht verkregen in de datakwaliteit? (N = 210)

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 7 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.

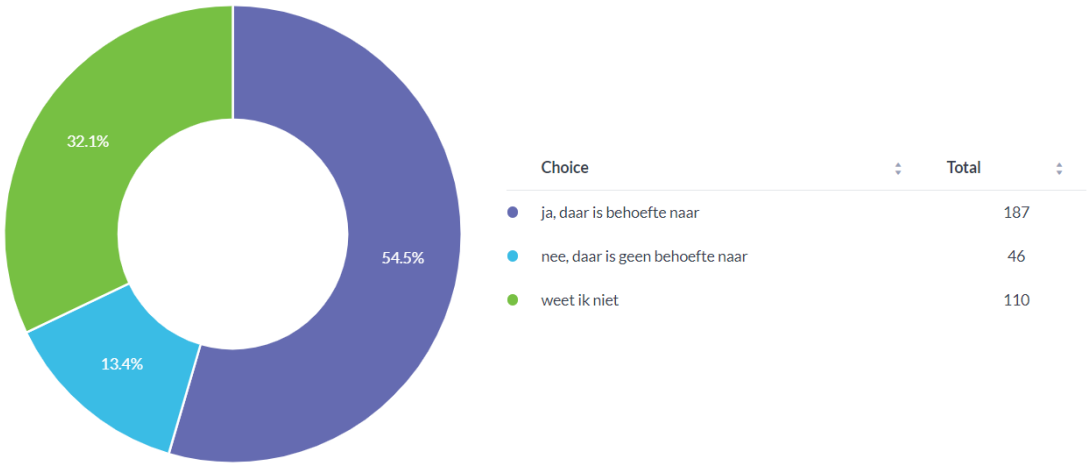
Q18. Welke aspecten worden (automatisch) getoetst? (N = 145)

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.



FIGUUR 8 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.

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%

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.

Conclusie

Allereerst laten de resultaten zien dat de respondenten op veel vlakken vrij gemixt zijn. Op het aantal uur per week, het niveau van expertise met betrekking tot data en de taken met betrekking tot data is veel variatie.

De enquêteresultaten laten daarnaast zien dat er grote belangstellig is voor datakwaliteit en dat er vaak problemen voorkomen door onvoldoende datakwaliteit. Het bewustzijn van respondenten ligt gemiddeld vrij hoog en veel respondenten voeren binnen de organisatie of het bedrijf al enige vorm van toetsing uit, met name het intern melden van fouten.

Op basis van de enquêteresultaten kan gesteld worden dat een dashboard voor het toetsen van datakwaliteit wel degelijk marktwaarde biedt.