Open(geo)data en datakwaliteit

Open data lijkt het nieuwe modewoord binnen ons (geo)ICT werkveld. Kortweg betekent het dat we beschikbare informatie vrijelijk beschikbaar stellen aan geïnteresseerde gebruikers. Door de aandacht hiervoor, ontstaat meer en meer een mind-set die het mogelijk maakt de potentiële voordelen ervan daadwerkelijk te gaan realiseren.
Dat het in de praktijk minder eenvoudig is dan het openen van de datasluizen, zal niet als een verrassing komen. Vragen zat: Wie mag wat zien, hoe zit het met de juridische aansprakelijkheid, wat is technisch nodig om de data te ontsluiten, of vragen over de rol van de overheid en die van het bedrijfsleven…. kortom voer voor een serie artikelen en nog meer seminars. In dit artikel wil ik met u naar een belangrijk aandachtspunt in dit kader kijken: de *Datakwaliteit.* Tevens naar de interessante rol die geo technologie kan spelen bij het realiseren van de juiste kwaliteit van onze (open)data. Een onderwerp van belang als we over open data praten, maar ook zonder die aanleiding cruciaal voor organisaties.

Beproefde benadering

Ik ben een vervent voorstander van het breed delen van informatie. Vanaf het prille begin in de vroege negentiger jaren, dat we informatie in geo-grafisch georiënteerde systemen gingen registreren en gebruiken, ben ik daarmee actief geweest. De meerwaarde van het benutten van een GIS was, en is, het op een in- en overzichtelijke wijze kunnen presenteren van grote hoeveelheden informatie. Een belangrijke stroomversnelling in het breed kunnen benutten van GIS zagen we rond de millenniumwisseling. Toen maakte de (standaard) hardware, de databases en niet te vergeten de opkomst van het internet het daadwerkelijk mogelijk grote groepen gebruikers te bedienen. Dit vormde destijds ook het fundament toen we Vicrea startten. Het begrip open data bestond als zodanig nog niet en informatie delen bleef veelal binnen één organisatie. Maar we hebben in die fase veel ervaring op kunnen doen met het breed delen van informatie en de bijbehorende aandachtspunten op het gebied van de datakwaliteit bij de hand gehad.

Meerwaarde data delen

Belangrijke drijfveren voor het breder delen van informatie liggen in basis in het streven elke medewerker *tijdig* de beschikking te geven over de *juiste* informatie om de gevraagde taak uit te kunnen voeren. Dan immers bereiken we een maximale efficiency binnen de processen, kunnen organisaties snel de juiste beslissingen nemen en daardoor effectief handelen. Voor zover dat nog niet voldoende motivatie is, zien organisaties zich door toenemende externe eisen genoodzaakt de informatievoorziening naar een hoger plan te tillen. Zo dwingen de basisregistraties gemeenten gebruik te gaan maken van centrale registraties (GBA, BAG, BGT,…) en eist de energiekamer een gedetailleerde verantwoording van de netbeheerders over hun assets en het gebruik ervan.
Vanuit onze geo-achtergrond is deze ontwikkeling extra interessant, omdat wij als geen ander in staat zijn om via de locatie component informatie bronnen de “open data” aan elkaar te “verbinden” en daarmee interessante kansen te creëren voor gebruikers.

Tijdig de juiste data

Sleutelwoorden in het voorgaande zijn de woorden *tijdig* en vooral *juiste*. Om dit belang te onderstrepen een praktisch voorbeeld: Indien een beperkt aantal medewerkers gebruik maakt van een gegevensbestand dat niet volledig juist (en dus betrouwbaar) is hoeft dit geen probleem te zijn. Men weet hiervan en zal in voorkomende gevallen bij elkaar te rade gaan of in een bepaalde situatie de gegevens correct zijn, en zo niet wat de feitelijke situatie is. Indien we diezelfde gegevens met een grote groep gaan delen is dit kwaliteitsbewustzijn over de tekortkomingen en de mogelijkheid de data te verifiëren bij collega's er niet. Het moge duidelijk zijn dat we in die situatie hogere eisen aan de data stellen. Het begrip *juiste* is te vertalen naar eisen die we aan de data stellen. Zo zal deze volledig moeten zijn. De werkelijkheid correct en actueel weergeven. Begrijpbaar voor de gebruiker moeten zijn en niet voor tweeërlei utleg vatbaar. Maar ook consistent in de zin dat we overal hetzelfde met een begrip bedoelen. Kortom de data moet betrouwbaar zijn. *Tijdig* heeft daarbij een meer technische invalshoek. Kan de gebruiker snel genoeg bij de informatie? Bij een meldkamer is het als voorbeeld onbestaanbaar dat op zich betrouwbare informatie pas na een paar minuten beschikbaar komt. Tijdig kunnen we ook vertalen naar de wijze van presenteren van de informatie. Ziet de gebruiker alle relevante informatie direct in één overzichtelijke presentatie, of moet hij/zij verschillende schermen / applicaties doorlopen …

Natuurlijke weerstand

Mede ingegeven door de open data wens om informatie te delen en de verhoogde (externe) eisen, investeren veel organisaties momenteel veel tijd en geld in datakwaliteit-trajecten. Een interessant fenomeen daarbij vormen de natuurlijke barrières die daarbij vaak te doorbreken zijn bij de eigenaren van de data. Voortbordurend op het eerdere voorbeeld is het voorstelbaar dat het team met de niet 100% betrouwbare dataset, naar de buitenwacht het beeld overeind houdt dat zij hun gegevens prima op orde hebben. Door de gegevens breder beschikbaar te gaan stellen, en in dat proces deze te relateren aan (of in dit geval beter, te confronteren met) andere gegevens valt dat positieve beeld vaak plots in duigen…Logisch dat er niet door iedereen van harte mee zal willen werken. Vanuit het belang van de organisatie geredeneerd, is deze wetenschap juist een extra motivatie om in te zetten op het verbinden van gegevensverzamelingen. Door de discrepanties die daarbij ontstaan inzichtelijk te maken is een belangrijke kwaliteitsslag te maken.

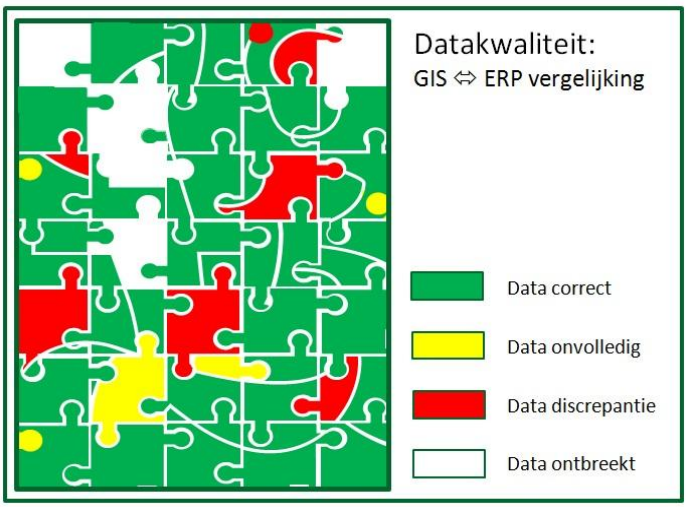
1a Mijn data is ok!

1b Data delen en onderzoeken

1c Data blijkt minder ok….Discrepanties wegwerken *1d Data is nu echt ok!*

Geo en datakwaliteit-trajecten

Geo technologie kan een belangrijke rol binnen trajecten om data te verbeteren spelen. In een kennissessie van Ruimteschepper bij Alliander kwamen we vanuit de vraag waar geo op dit moment een significante bijdrage voor beheerders van infrastructuren kan leveren, vrij snel op de ondersteuning van kwaliteitsverbetertrajecten. Daarbij niet alleen kijkend naar de kwaliteit van de geografische data, maar breder naar alle informatie over de “assets” van de organisatie. Door bijvoorbeeld verschillende gegevensbestanden op basis van de locatiecomponent (assetcode, adres, x/y, perceel,..) bij elkaar te brengen, zijn via thematische geo-presentaties de discrepanties inzichtelijk te maken. Denk aan “niet gevulde velden”, “verschillende waarden” ed. Alliander gebruikt dit o.a. om de voortgang van haar datakwaliteits project te monitoren. Vanuit genoemde sessie vertalen geo-experts deze aspecten momenteel naar een Ruimteschepper workshop om betrokkenen binnen organisaties bekend te maken met dergelijke mogelijkheden. Daarmee is een belangrijke versnelling in deze veelal omvangrijke projecten te realiseren.



1 Geopresentatie t.b.v. datakwaliteittraject

Is goed, goed genoeg?

Ten slotte nog een vaak terugkerende vraag rondom het beschikbaar stellen van data in relatie tot de datakwaliteit. “Wanneer is het verantwoord om data beschikbaar te stellen?”. Moet deze dan 100% juist zijn voor alle aspecten? In dat geval is de volgende definitie nuttig:

"Datakwaliteit is de mate waarin data geschikt is voor het doel waarvoor ze gebruikt wordt".

Dat betekent dus dat niet in alle gevallen de 100% eis nodig is. Wel dat we bewust moeten zijn van het voorziene gebruik van de data. Om dit te illustreren een voorbeeld uit een recent project waar ik bij betrokken was: Binnen de doelstelling informatie organisatiebreed beschikbaar te stellen, is ten aanzien van de vraag welke gegevens hier wel en welke hier niet voor in aanmerking kwamen een

duidelijke keuze gemaakt. Besloten is data, tenzij er evidente veiligheidsrisico's aan kleefden, beschikbaar te stellen. Daarbij is naar de organisatie nadrukkelijk aangegeven dat de data de "as-is" data uit de beheerbestanden betrof en niet gegarandeerd 100% correct was. Dit gekoppeld aan de vraag om bij geconstateerde fouten deze te melden. Resultaat was dat vrij snel fouten in bijvoorbeeld de registratie van het netwerk, die al jaren in de systemen bleken te zitten, gemeld werden en snel gecorrigeerd zijn. Zo heeft de geo-gebaseerde intranet raadpleegomgeving als een belangrijke katalysator voor het verbeteren van de datakwaliteit gefungeerd. Daarmee een goed voorbeeld hoe geo technologie effectief is in te zetten in deze context.

Samenvattend, datakwaliteit is een belangrijk aandachtspunt om de potentiële meerwaarde van open data te kunnen benutten. Geo-technologie biedt daarbij unieke kansen om de gewenste kwaliteit te kunnen realiseren!

j.roodzand@ruimteschepper.nl
j.roodzand@net4s.nl
@JanRoodzand