Big data klaar voor gebruik?
De coördinatie van de dataketen

Janssen, Marijn; van der Voort, Haiko

DOI
10.5553/Bk/092733872016025001003

Publication date
2016

Document Version
Final published version

Published in
Bestuurskunde

Citation (APA)
Janssen, M., & van der Voort, H. (2016). Big data klaar voor gebruik?: De coördinatie van de dataketen. Bestuurskunde, 25(1). DOI: 10.5553/Bk/092733872016025001003

Big data klaar voor gebruik?

De coördinatie van de dataketen*

Marijn Janssen & Haiko van der Voort

In het proces van creëren, opschonen, verwerken en gebruik van data zijn verschillende partijen betrokken, met verschillende deskundigheid, motieven en belangen. Dit maakt overdracht van kennis belangrijk, maar ook moeilijk. Bedoelen partijen hetzelfde? Begrijpen partijen hetzelfde? Willen partijen hetzelfde? Om de waarde van data te kunnen bepalen moet het gehele onderliggende proces worden geanalyseerd en begrepen. Een belangrijk deel van dit proces kan buiten de eigen organisatie liggen. De uitdaging voor overheden als gebruikers van big data is de datakwaliteit te managen, om zo de grondslag van hun beslissingen stevig te houden.

Een onthechting van de mens?

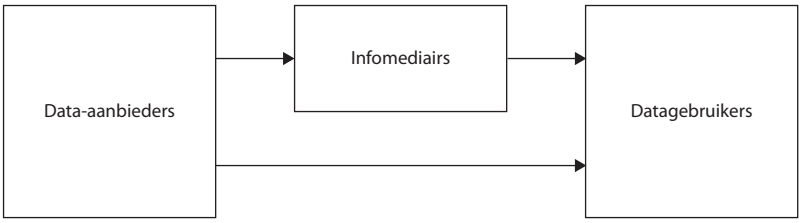
Voordat u het weet, genereert u data. Uw smartphone kan uw locatie meten, uw beweging vaststellen, net als de luchtdruk en uw hartslag. Uw auto rijdt langs verkeerslussen en uw afspraak met de gemeente wordt geregistreerd. Bij het browsen laten mensen impliciet hun voorkeuren en interesses achter door naar bepaalde pagina's te gaan, op 'like' te klikken, een 'tweet' te versturen maar ook expliciet door opmerkingen te plaatsen op discussiefora. In de interactie met een organisatie kan ook veel data verzameld worden over de kwaliteit van de ingevoerde informatie, over de persoonlijke voorkeuren, de (sociale) problemen waarmee iemand worstelt. Dit zijn allemaal voorbeelden van plaatsen waar grote hoeveelheden data gegenereerd worden.
Het verzamelen, verwerken en gebruiken van deze grote hoeveelheden data wordt vaak aangeduid als 'big data'. De waarde van big data komt uit het combineren van verschillende bronnen om nieuwe inzichten te creëren en verborgen waarde te ontsluiten (Janssen, Estevez, & Janowski, 2014). Zo kan big data zorgen voor betere benutting van schaarse bronnen en voor op maat gesneden dienstverlening, ook door overheden (Chen & Hsieh, 2014).

Mensen lijken hier een steeds kleinere rol te spelen. Het zijn computers die de data verwerken en soms combineren. Computers zijn ook beter in staat dan mensen om uit een grote hoeveelheid data interessante, betekenisvolle patronen te vinden. De beloftes van big data gaan vaak gepaard met een impliciete aanname dat big data processen zich onthechten van mensen. Het begrip *internet of things* – vaak in één adem genoemd met big data – geeft het al aan: alledaagse voorwerpen kunnen zonder inmenging van de mens gegevens genereren en uitwisselen.

* Prof. dr. ir. M.F.W.H.A. Janssen is hoogleraar aan de Technische Universiteit Delft. Dr. H.G. (Haiko) van der Voort is universitair docent aan de Technische Universiteit Delft.

Bestuurskunde 2016 (25) 1
doi: 10.5553/Bk/092733872016025001003

Figuur 1 Schematisch overzicht van de actoren



Zelfs *the end of theory* is al aangekondigd (Anderson, 2008). In plaats van de inductieve worsteling van sociale wetenschappers kunnen computers ons verder helpen met automatische deductie.

Maar speelt de mens een kleinere rol bij big data als grondslag voor beleidskennis? Hiervoor is behoefte aan het conceptualiseren van de big data keten van datageneratie tot aan datagebruik, bij wijze van spreken van verkeerslus naar mobiliteitsbeleid. Gaat het automatisch? Gaat het vanzelf? Op deze manier benaderen we big data als een coördinatievraagstuk in een dataketen en vragen we ons hardop af wat er van de mens nodig is voordat big data klaar zijn voor gebruik.

Een procesblik op big data: van ruwe data naar gebruik

Big data wordt vaak als een technisch fenomeen beschreven. Ten onrechte, want er zijn ook veel typisch organisatorische vraagstukken. In een big data proces worden ruwe data gecreëerd, opgeschoond, verrijkt, gecombineerd en uiteindelijk gebruikt. Figuur 1 geeft dit proces schematisch weer. De actoren die hierin een rol spelen, lijken op die van ieder andere keten. Zo zijn er dataleveranciers en datagebruikers. Tussen leveranciers en gebruikers bevinden zich veelal infomediairs die data uit verschillende bronnen kunnen combineren en integreren, verrijken of zelfs al complete analyses uitvoeren voor gebruikers. Zo ontstaat het beeld van variëteit: er zijn veel verschillende actoren die zijn betrokken bij verschillende delen van dit proces, veelal zelfs binnen een enkele organisatie. Bovendien kunnen we met een keten twee soorten coördinatieopgaven onderscheiden, namelijk coördinatie binnen een organisatie en coördinatie tussen organisaties.

Eerste coördinatieopgave: complexiteit binnen organisaties

Variëteit geldt niet alleen voor de betrokken actoren. Verzamelaars van data worden geconfronteerd met veel verschillende databronnen. Sommige bronnen zijn gestructureerd, anderen niet. Sommige data zijn beschikbaar voor iedereen, andere niet. Sommige data zijn lang houdbaar, andere niet. Data over criminaliteit en hotspots van afgelopen maand hoeven bijvoorbeeld deze maand niet meer

geldig te zijn. Bij het weer zijn we ons hiervan bewust, maar bij andere data niet altijd.

In een volgende stap worden data veelal geïntegreerd. Dit is waar data-analisten hun rol hebben. Wat gebeurt er als data die zijn verkregen op verschillende tijden, gecombineerd worden? Of data die in Groningen zijn verzameld over de economische groei, worden gecombineerd met gebruikersdata uit Maastricht? Wat als een gebruikersprofiel wordt opgesteld van iemand met een veelvoorkomende naam en de databronnen van verschillende personen zijn?

In een daaropvolgende stap worden beschrijvende en voorschrijvende analytics en statistieken toegepast en gevisualiseerd. Dit vereist diepgaande kennis over het mogen gebruiken van deze methoden en aannames die gedaan worden. Zo ontstaat er informatiescheefheid binnen het proces, met daarbij mogelijk een (extra) kennisvoorsprong van analisten. Met andere woorden: data-analyse als vernieuwde tak van sport leidt tot nieuwe verhoudingen tussen analisten en managers binnen een schakel.

De drie stappen zijn uiteraard een simplificatie van het big data proces. De stappen herbergen op zichzelf al technische en sociale complexiteit. Het plaatje wordt echter nog complexer met de gedachte dat de stappen elkaar beïnvloeden. Bijvoorbeeld: motieven achter dataverzameling kunnen verschillen. Dit maakt vervolgens het combineren van data lastig. Ruwe data worden vaak opgeschoond om fouten eruit te halen, om datakwaliteit te verbeteren en om verwerking van grote hoeveelheden data te vergemakkelijken. Dit is echter risicovol voor actoren verderop in het proces, omdat juist de uitzonderingen kunnen leiden tot de inzichten uit de data, afhankelijk van het motief van de gebruiker. Zo kan bij het analyseren van digitale interactie met de overheid zo'n uitzondering erop duiden dat iets toch niet voor iedereen duidelijk is. Door het helderder formuleren van een vraag kan dit worden voorkomen.

Kan ervan worden uitgegaan dat de verschillende data ook zonder meer voor een ander doel gebruikt kunnen worden? Zijn dan wel de juiste meetmomenten en -instrumenten genomen? Zijn de data al eens eerder geïnterpreteerd en welke assumpties lagen hieraan ten grondslag?
Dit soort vragen zijn door een enkele persoon wellicht goed te beantwoorden, maar wanneer data door vele organisaties of afdelingen worden verzameld, dan is beantwoording een kwestie van coördinatie.

Tweede coördinatieopgave: transfers tussen organisaties

Het voorgaande geeft aan dat een procesblik op big data verschillende coördinatievraagstukken blootlegt. Over iedere schakel van de keten van generatie tot gebruik bestaan er lastige overdrachten van informatie. Bij deze overdrachten bestaan er noodzakelijkerwijs aannames over de data, over het dataproces en over de actoren in de andere schakels. Met name deze overdrachten, of transfers, zijn

Bestuurskunde 2016 (25) 1
doi: 10.5553/Bk/092733872016025001003

gevoelig voor fouten en misverstanden. We gaan nu dieper op deze transfers in en onderscheiden daarbij drie organisatorische complexiteiten, van operationeel tot bestuurlijk.

Semantiek: bedoelen we hetzelfde?
Voor het gebruik van data moet men kennis hebben van de betekenis van data, oftewel de semantiek. Kennis van semantiek is nodig om data te interpreteren, hoe deze al bewerkt is, wat de kwaliteit is en uiteindelijk wat de data waard zijn. De betekenis van criminaliteitsdata zal bijvoorbeeld per persoon verschillen. Als uw fiets net gestolen is, denkt u wellicht dat criminaliteitsdata over gestolen fietsen gaan; de meeste criminaliteitsdata gaan echter over inbraken, berovingen et cetera. Maar hoe kan men weten of de data alle criminaliteit omvatten? Immers, veel lichte misdaden worden niet gemeld en geregistreerd. Bovendien is het belangrijk meer te weten over de authenticiteit van de bron. Zijn de data verzameld door de politie (wat registreren deze wel en niet?) of door de slachtoffers? Weten we dat als we het over 'gestolen fiets in Amsterdam' hebben of het over hetzelfde incident gaat? Semantiek is essentieel voor het voorkomen van misverstanden. Er zijn aldus eenduidige en geaccepteerde categorieën van fenomenen nodig waaraan mensen dezelfde betekenis geven om de kwaliteit van transfers te bevorderen. Dit impliceert echter dat ieder fenomeen een breed geaccepteerd oormerk moet krijgen. En dat impliceert een enorme standaardisatie-opgave. Wellicht is deze opgave per project te overzien, maar de ambities van big data reiken hoger.

Kennis: begrijpen we hetzelfde?
Voor iedere procesfase is een ander soort deskundigheid vereist. Voor het verzamelen van data zijn mensen nodig met kennis over technische apparaten, bijvoorbeeld om kentekens te registreren met camera's of luchtvervuiling te meten, en hoe er met deze apparaten omgegaan moet worden. Ook zijn er softwareontwikkelaars nodig, die zorgen dat mensen met elkaar kunnen discussiëren. In een daaropvolgende schakel zijn meer database-experts te vinden, die zorgen dat data opgeschoond kunnen worden. Dit opschonen wordt al snel als 'technisch' gezien. In de procesfase daarna gaan de mensen met statistische en *analytics* kennis aan de slag om de data te analyseren en er conclusies uit te trekken. Daar waar de voorgaande mensen vaak geen domeinkennis hebben, moeten deze *data scientists* samenwerken met domeinexperts om chocola te maken van de data. Wat betekent het dat criminaliteit op een bepaalde plaats meer voorkomt? Zonder domeinexperts kunnen verkeerde conclusies worden getrokken, vergelijkbaar met het bekende voorbeeld dat 'vrouwen die aan de kust wonen vaker zwanger zijn'. Deze analyse geeft weer dat er veel transfers tussen disciplines nodig zullen zijn. De gevolgen van gebrekkige communicatie tussen disciplines laten zich raden: assumpties achter data blijven impliciet, *disclaimers* komen niet door en wat zacht is wordt hard in de volgende schakel.

Strategisch gedrag: willen we hetzelfde?

De stap van informatiescheefheid naar strategisch gedrag is klein. Hebben partijen in het proces om van een eventuele kennisvoorsprong gebruik te maken? Dit kan zowel in actieve zin (bewust data bewerken of presenteren in lijn met een eigen belang) als in passieve zin (slonzigheid uit efficiencyoverwegingen). Kernvraag is hier wie belang heeft bij hoge datakwaliteit. Leveranciers en analisten kunnen een economisch belang hebben om te leveren. Ze kunnen bijvoorbeeld alleen de data doorgeven die in hun beeld past en welke gunstig voor hen zijn. Facebook speelt bijvoorbeeld films af als u uw Facebookpagina aan het bekijken bent. Ze tellen de film mee als 'bekeken' als deze drie seconden heeft gespeeld, maar mensen kunnen iets anders aan het bekijken zijn om vervolgens weer verder te scrollen zonder dat ze de film hebben bekeken. Adverteerders die betaalden per keer dat de film bekeken was, waren hier niet blij mee. Bij het opschonen van de data kan een vergelijkbaar proces plaatsvinden. Aan de gebruikerskant kan het kwaliteitsbelang worden gecompromitteerd door een (te) hoge druk om goede prestaties kenbaar te maken. Hoe speculatief het bovenstaande ook moge klinken, coördinatie zal ook bij big data een politiek proces zijn, met daarbij een grote rol voor bestuurlijk overleg en soms verificatie door derden.

Conclusies: Big data als coördinatie-opgave

Big data en *internet of things* claimen een vergaande onafhankelijkheid van de mens maar tegelijkertijd voeren mensen stappen uit en moeten ze hun activiteiten op elkaar afstemmen. Er is geen depolitisering en strategisch gedrag blijft een rol spelen. Het benutten van big data vergt intensieve coördinatie tussen mensen en tussen organisaties. Voor ieder initiatief is een analyse van het dataproces en de participerende actoren verplicht huiswerk. Zonder een dergelijke analyse zal de gebruiker onvoldoende kennis hebben van de processen die zich afspelen en wat de valkuilen zijn. Bovendien zullen de kosten van kwaliteit worden onderschat. Die kosten gaan niet alleen over de data zelf, maar veeleer over de transacties tussen actoren en de benodigde maatregelen om deze transacties te faciliteren, zoals standaardisatie en verificatie, maar vooral veel overleg.

Literatuur

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*. Retrieved from http://www.wired.com/2008/06/pb-theory/

Chen, Y.-C., & Hsieh, T.-C. (2014). Big data for digital government: Opportunities, challenges, and strategies. *International Journal of Public Administration in the Digital Age, 1*(1), 1-14.

Janssen, M., Estevez, E., & Janowski, T. (2014). Interoperability in big, open, and linked data--organizational maturity, capabilities, and data portfolios. *Computer, 47*(10), 44-49. doi: 10.1109/MC.2014.290

Bestuurskunde 2016 (25) 1
doi: 10.5553/Bk/092733872016025001003