Onbegonnen werk of toch niet zo anders?

Kwaliteit van big data

Big data bieden volgens tal van wetenschappers en beleidsmakers veel kansen en bedreigingen. Zo zouden big data inzicht kunnen bieden in verborgen armoede, helpen met het ontdekken van zeldzame ziektes, maar ook onze gemeenteraadsverkiezingen beïnvloeden.[1] 'Big data' klinkt als een belofte en een bedreiging voor de maatschappij. Is dat ook zo? Is dit meer dan slechts een hype? En zijn big data werkelijk anders dan andere data?

TEKST **DONNA PLUGGE EN NADINE GROFFEN***

Als studenten van de master Archiefwetenschap hebben wij een verkennend onderzoek gedaan naar hoe bedrijven de kwaliteit van data beoordelen en het uiteindelijk ook waarborgen. Hiervoor hebben we onder andere de datakwaliteitsstandaard ISO/IEC 25012 bestudeerd en zijn we in gesprek gegaan met twee bedrijven: Guidion en Data Kitchen. De focus van de interviews lag op het achterhalen hoe de bedrijven bepalen wat kwalitatieve data zijn en hoe ze de kwaliteit waarborgen binnen hun processen. Oftewel: op welke manier definiëren en herkennen deze twee bedrijven kwaliteit van data, en welke overeenkomsten zijn hieruit af te leiden die ook voor andere bedrijven en overheden interessant zijn

Onderbelicht: de kwaliteit van big data

In het tijdperk van big data wordt informatie op grote schaal verzameld, vastgelegd en beheerd. Er is geen algemeen geaccepteerde definitie voor big data. Het is een containerbegrip geworden, maar het bestaat in ieder geval uit: 'gestructureerde en ongestructureerde datasets die niet kunnen worden opgeslagen en verwerkt door normale softwaretools, zoals relationele databases'.[2] Daarnaast onderscheiden big data zich ten opzichte van 'gewone' data als het gaat om veel grotere volumes informatie, met daarin veel diversiteit en een snelle veranderende verwerkingssnelheid. Maar het besef groeit steeds meer dat er nog een belangrijk, en tot nu toe enigszins onderbelicht, aspect is aan big data: de kwaliteit ervan.[3] Want voor big

Data spelen een grote rol in de besluitvorming van een bedrijf

data geldt hetzelfde als voor 'gewone' informatie: je hebt er pas iets aan als je van de kwaliteit op aan kunt. Maar wat is datakwaliteit dan eigenlijk? Hoe kun je het herkennen, en nog belangrijker, borgen? En waarom is datakwaliteit eigenlijk zo belangrijk om te onderzoeken?

Slechte kwaliteit van data kan grote gevolgen hebben. Denk aan missende links of gegevens, onbetrouwbare data en verkeerde representatie van data.[4] Bedrijven nemen vaak grootschalige beslissingen op basis van hun data. Slechte

Foto: Vintage Tone | Shutterstock.com

kwaliteit van data kan bedrijven problemen bezorgen, zoals een lagere omzet of ontevreden klanten.[5] Ook overheden hebben belang bij het waarborgen van datakwaliteit om hun primaire taken te kunnen uitvoeren. Er wordt wel eens gezegd dat data het nieuwe goud zijn.[6] Data spelen dus een grote rol in de besluitvorming van een bedrijf. Genoeg reden om een kijkje te nemen in de manier waarop bedrijven met de kwaliteit van hun data omgaan.

Guidion

Guidion levert servicemonteurs voor telecom- en energiebedrijven. Guidion maakt namens deze bedrijven afspraken met de klanten en stuurt servicemonteurs op pad.[7] Het bedrijf is dus een intermediair in de dienstverlening en het centrale punt waar alle data van zowel klanten als opdrachtgevers bij elkaar komen. Guidion beheert niet alleen informatie aangaande lopende aanvragen en opdrachten, maar houdt ook feedbackgegevens van klanten bij over de service van de monteurs. Die informatie komt ongestructureerd binnen, dat wil zeggen dat klanten op verschillende wijze feedback geven. De klanten sturen zowel foto's en video's als cijfers of teksten. Deze feedback-data beschouwt Guidion als big data, omdat ze ongestructureerd zijn. De rest van de data die zij beheren is echter gestructureerd. Dat wil zeggen dat de data zich bevinden in relationele databases met gestructureerde invoervelden. In de interviews bleek dat het begrip big data verschillend wordt gebruikt. Dat een bedrijf met veel data werkt, wil dus niet meteen zeggen dat het ook om big data gaat. De kwaliteit van de data die Guidion beheert, blijft bepalend voor zowel het soepel lopen van de dienstverlening van de telecom- en energiebedrijven, als voor de kwaliteitsborging van hun service.

Maakt het eigenlijk verschil voor de kwaliteitsborging of data 'gewoon' of 'big' zijn?

Voor de beoordeling van de kwaliteit van zijn data gebruikt Guidion drie kwaliteitsdimensies: correctheid, beschikbaarheid en integriteit. Guidion beschouwt data correct als ze inhoudelijk kloppen, zoals de juiste weergave van een klantadres. De beschikbaarheid gaat zowel over de beveiliging van data (wie mag waar bij?) als over de technische toegankelijkheid. Onder integriteit verstaat Guidion dat de data consistent worden ingevuld. Bij het invoerveld huisnummer mogen medewerkers bijvoorbeeld alleen cijfers invoeren (1) en geen letters (één).

Guidion krijgt dus van verschillende partijen grote datasets aangeleverd. Guidion kan niet sturen op de kwaliteit van de aanlevering van deze datasets, maar wel op wat ze zelf doen. Binnen het bedrijf wordt datakwaliteit gewaarborgd. Zo gebruiken ze bijvoorbeeld standaardinvoervelden om data consistent vast te leggen in het systeem en trainen ze de eigen medewerkers om uniform te werken. Desondanks blijkt correctheid van data het moeilijkste aspect om te bewaken. Correctheid is afhankelijk van het juist invullen van data in voornamelijk vrije invoervelden. Menselijke fouten zijn hier vaak snel gemaakt, ondanks de vele trainingen die medewerkers krijgen.

Foto: Jirapong Manustrong | Shutterstock.nl

Data Kitchen

Data Kitchen benadert datakwaliteit op een andere manier, omdat het als sales- en marketingbedrijf andere bedrijven adviseert om de kwaliteit van hun interne data te verbeteren.[8] Interne data worden door Data Kitchen *fit for use* gemaakt, oftewel systemen worden 'geschoond' van foute en dubbele data. Data Kitchen helpt andere bedrijven bijvoorbeeld met het maken van klantenprofielen met de juiste persoonsgegevens, zoals naam, adres en telefoonnummer. Sommige klanten van Data Kitchen beheren daadwerkelijk big data, maar de meeste bedrijven maken net zoals Guidion gebruik van 'gewone' data uit gestructureerde systemen. Maar maakt het eigenlijk verschil voor de kwaliteitsborging of data 'gewoon' of 'big' zijn?

Data Kitchen beoordeelt de kwaliteit van de data via het *ACCU*-principe. Daarmee wordt gekeken of de data actueel, compleet, correct en uniek zijn. Data zijn actueel als ze de werkelijkheid weerspiegelen. Dit betekent dat de data up-to-date moeten zijn, zoals de adreswijzigingen die doorgevoerd moeten worden. Verder zijn data compleet als ze alle benodigde informatie bevatten om de vraag van een bedrijf te beantwoorden. Een naam en een telefoonnummer kan bijvoorbeeld voldoende zijn voor een klantenprofiel als het bedrijf de adressen niet nodig heeft. De correctheid staat voor de juistheid van data. De correctheid van data wordt eigenlijk bepaald door de andere drie kwaliteitsdimensies. Data zijn correct als ze recent en compleet zijn, en geen dubbelen bevatten. De kwaliteitsdimensie uniekheid is verbonden aan dubbele data, waarbij een data-element maar één keer voor mag komen.

Data Kitchen waarborgt de kwaliteit van data met datagovernance. Dit betekent plannen, controleren en toezicht houden over datamanagement en het gebruik van interne en externe data.[9] Allereerst worden de datastromen bin-

nen een of meerdere systemen in kaart gebracht. De systemen en processen moeten op de juiste manier ingericht worden om datakwaliteit te kunnen waarborgen. Het is bijvoorbeeld belangrijk dat werknemers van verschillende afdelingen dezelfde 'taal' leren spreken en onder bijvoorbeeld het woord 'klant' allemaal hetzelfde verstaan. Data Kitchen maakt echter geen onderscheid in kwaliteitsdimensies voor big of small data. Kwaliteitswaarborging van data is daarentegen wél afhankelijk van een andere factor, namelijk die van de klantvraag. Soms willen klanten een eenmalige opschoning van hun data en lopen daarmee het risico dat de data na verloop van tijd niet meer actueel, compleet, correct of uniek zijn.

Vergelijking van de bedrijven en ISO/IEC 25012

Beide bedrijven geven aan dat ze datakwaliteit belangrijk vinden, maar werken niet met een internationale kwaliteitsstandaard, zoals de ISO/IEC 25012. Zo heeft Data Kitchen zelf de ACCU-standaard ontwikkeld op basis van literatuurstudie en hun eigen praktijkervaring. Ondanks dat beide bedrijven aangeven er niet mee te werken, hebben wij ze wel vergeleken met de ISO/IEC 25012. In ISO/IEC 25012 worden verschillende kwaliteitsdimensies benoemd. Het blijkt dat alle drie de kwaliteitsdimensies die Guidion hanteert het meest aansluiten bij de ISO/IEC 25012. Bijvoorbeeld de dimensie beschikbaarheid. De ISO/IEC 25012 en Guidion geven dezelfde betekenis aan deze dimensie. Beiden hebben het over de beveiliging van de data. Maar wat zegt dit eigenlijk?

Wat datakwaliteit is, verschilt sterk per organisatie en het doel dat bereikt moet worden

Dat de bedrijven verschillend met datakwaliteit omgaan is niet zo vreemd. Data Kitchen en Guidion hebben een heel andere invalshoek. Dit komt mede doordat beide bedrijven de data anders gebruiken in het primaire proces van hun organisatie. Data Kitchen geeft advies aan bedrijven om hun data bruikbaar te maken. Guidion heeft data nodig om zelf goed te kunnen functioneren. Daarom kan kwaliteit van data ook niet met één kwaliteitsstandaard beoordeeld worden. Een kwaliteitsstandaard kan bedrijven of overheden wel een kader bieden om datakwaliteit te borgen.

Conclusie

De bedrijven Guidion en Data Kitchen zijn zich ervan bewust dat de kwaliteit van data erg belangrijk is. Wat datakwaliteit is, verschilt echter sterk per organisatie en het doel dat bereikt moet worden. Datakwaliteit kan in de praktijk geborgd worden door vaste keuzemogelijkheden in systemen in te bouwen, trainingen aan te bieden aan werknemers of een datagovernancestrategie te bepalen. Maar voor Data Kitchen bepaalt uiteindelijk de klant de mate van kwaliteit, omdat het bedrijf afhankelijk is van hun prioriteiten, mensen en middelen. Guidion is daarentegen afhankelijk van zijn eigen data om het primaire proces goed uit te kunnen voeren. Al met al is er voor de bedrijven Guidion en Data Kitchen geen verschil tussen de beoordeling en waarborging van big data of andere soorten data. Alle data wordt aan de hand van dezelfde kwaliteitseisen beoordeeld. Of het nu om big data of small data gaat, data zijn gewoon data. Datakwaliteit is echter niet 'gewoon', maar vergt per organisatie, per proces en per werknemer veel aandacht. ●

* Met dank aan Natascha Gadellaa en Suzi Szabó.

Noten

1 https://www.platform31.nl/nieuws/big-data-voor-inzicht-in-verborgen-armoede; https://www.icthealth.nl/nieuws/big-data-help-bij-vaststellenn-lymfeklierkanker-bij-vrouwen-met-borstimplantaten/; https://www.bnr.nl/nieuws/politiek/10339427/gaat-big-data-de-gemeenteraadsverkiezingen-bepalen (websites geraadpleegd op 6 februari 2018).

2 Donatella Firmani, Massimo Mecella, Monica Scannapieco, Carlo Batini, 'On the Meaningfulness of 'Big Data Quality' (Invited Paper)', Data Science and Engineering 1:1 (2016) 6-20, aldaar 13.

3 Ibidem, 13.

4 Viktor Mayer-Schönberger en Kenneth Cukier, Big data: A Revolution That Transforms How We Live, Work, and Think (Amsterdam 2013) 54 en 56.

5 Hazen, B. T., Boone, C. A., Ezell, J. D., en Jones-Farmer, L. A., 'Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications', International Journal of Production Economics 154 (2014) 1-28, aldaar 3. Wang, R. Y., en Strong, D. M., 'Beyond accuracy: What data quality means to data consumers', Journal of management information systems, 12(4) (1996) 5-33, aldaar 7.

6 Harari, Yuval Noah., Homo Deus: A brief history of tomorrow (Random House 2016).

7 https://www.guidion.com/

8 http://www.datakitchen.nl/

9 P. Cupoli de e.a., DAMA-DMBOK2 Framework (DAMA International 2014) 10.

Donna Plugge
donna.louise@live.nl

Donna Plugge is tweedejaarsstudent master Archiefwetenschap en werkt als archiefmedewerker bij het Teylers museum.

Nadine Groffen
Nadine.Groffen@nationaalarchief.nl

Nadine Groffen is tweedejaarsstudent master Archiefwetenschap en werkt als medewerker Invoer Actorenregister bij het Nationaal Archief.

Zie ook www.od-online.nl