

Data integriteit en ontketend denken

René Veldwijk

Managing Partner Ockham Groep BV, Vianen

Introductie spreker en onderwerp

- René Jan Veldwijk, 1961 te Bussum
- Econoom en informatiekundige, gegevensexpert, researcher en columnschrijver
- Opgegroeid in groot software huis (Raet) en sinds 1995 verbonden aan succesvolle ICT maatschap (Ockham Groep)
- Werkzaam voor private bedrijven en overheden. In 2007/2008 *finest hour* als ontwerper/bouwer van de grootste publieke administratie, de polisadministratie van UWV.
- Sinds oktober 2011 publiek gegaan met de stelling dat in de publieke ICT mislukken de norm is. Druppel voor parlementair onderzoek naar falende ICT bij overheden. Zie website **ockham.nl**.

En...toch gevraagd voor dit evenement!

Agenda voor vandaag

1. Data integriteit in enge en in ruime zin
2. De trieste *state of the art*.
3. Innovatie: de UWV Polisadministratie casus
4. Denken bij de overheid: autonomie versus integratie
5. De Amsterdamsche Spookburger casus: een ommekeer?
6. Naar een integere publieke gegevenshuishouding: de Trias Informatica
7. Afsluiting

1. Data integriteit in enge en in ruime zin

Theorie in het superkort:

1. Een database is een verzameling gegevens die is onderworpen aan strenge regels (constraints) die onbestaanbare (combinaties van) gegevens niet accepteren.
2. Voorbeelden: field formats, not nulls, domeinregels, PKs, AKs en user-defined constraints van arbitraire complexiteit.
3. Regels zijn bedoeld om te helpen garanderen dat de inhoud van een database correspondeert met de relevante werkelijkheid buiten het systeem.
4. Een database waarin geen enkele regel wordt geschonden is **integer in enge zin**.
5. Een niet integere database is per definitie geen adequate representatie van de werkelijkheid, maar het omgekeerde geldt niet.
6. Een database die de werkelijkheid goed representeert is **integer in ruime zin** (en dus ook in enge zin).
7. We willen integriteit in enge zin om integriteit in ruime zin te bereiken.

2. De trieste state of the art

ICT hulpmiddelen

1. Jaren 50: “Records” dwingen integriteit af. De rest doet de programmeur
2. Jaren 60/70: Datatypen en prerelationele DBMSen (sleutels)
3. Jaren 80: Theorie voorwaarts (RDBMSen, modelgedreven werken, Nolan groeimodel) implementaties stap terug.
4. Jaren 90/00: Eenvoudige regels afgedwongen door DBMS, moeilijke niet (ondanks DB Triggers)

ICT praktijk (gemiddeld. niet alleen overheid)

1. Data management is iets technisch
2. Migraties vanuit legacy en verlies aan focus op integratie van gegevenshuishoudingen leiden tot “kaartenbak+ databases”
3. Integriteitsregels die niet worden afgedwongen op DBMS niveau worden niet bewaakt.

3. Innovatie: de UWV polisadministratie casus

Ockham / High~T

1. Ockham (mijn bedrijf) focust zich – heel ouderwets – op modelgedreven systeemontwikkeling met behulp van een eigen dictionary-gedreven ontwikkelomgeving (High~T).
2. High~T bevat het complete gegevensmodel van een applicatie (dus ook alle integriteitsregels) en valideert mutaties.
3. Speciaal voor corrupte databases valideert High~T ook *on the fly*, dat wil zeggen bij het presenteren van data op any scherm.

UWV (en BD) / Polisadministratie

1. In 2007/8 gebouwd door Ockham na fiasco in 2006/7 (remember toeslagen ellende? mw. Thunnissen?)
2. Deelprobleem was grootschalige datacorruptie door niet controleren van LA-berichten door de BD
3. Opgelost met High~T (z.o.z.)

3. Innovatie: de UWV polisadministratie casus

PERSONENDASH_BSN - Personendashboard (BSN-ingave)

Ingave burgerservicenummer (BSN_INGAVE_DS) (1)

Burgerservicenummer 940000003 K Jong Ding

Health check button

Overzicht Inkomstenverhoudingen (LA_IKVD1) (1)

Sofinr	Naam	IKV Id	LHnr	Naam Adm.Eenh.	NumIV	Pers. nr	Sexe	DatAanvTvk	D.
1	940000003	Jong Ding, K	303733	123456789L01	Reservoir Dogs Restaurant	1	6027635	2	

Violations:
- R:Lopend fraudeonderzoek.

Inkomstenperiode AVB (0) Inkomsten-en-perioden Signaleringen Logging Trans
Historie (0) Inkomstenopgave Sector-Ris

Naamhistorie (2) Adreshistorie Geldigheidshistorie Sleutelhistorie

Inkomstenverhouding geldigheid historie (LA_IKVHISTD4) (6)

	DatAanv	Dat tot	RdnExFlex	Ttime in	Ttime eind
1	20060814	20070423	3	11-11-2007 11:11:11,666666	07-03-2008 04:01:04,266000
2	20060814	20070101	3	07-03-2008 04:01:04,266000	31-12-9999 00:00:00,000000
3	20070101	20070129		07-03-2008 04:01:04,266000	31-12-9999 00:00:00,000000
4	20070129	20070423	3		
5	20070521	20070618	1		
6	20070618	99991231	3	11-11-2007 11:11:11,666666	31-12-9999 00:00:00,000000

Violations:
- C:Constraint HIS_IKV_IKGH03 violated: Er mag geen overlap tussen de transactietijden in de historie zijn.

Violations:
- R:Code reden einde flexwerker ongeldig op datum aanvang.
- R:Code voor reden einde IKV flexwerker onbekend.
- R:Code einde inkomstenverhouding flexwerker alleen gevuld voor flexwerker en einddatum gevuld (ex 0 en 00)
- R:Code einde inkomstenverhouding flexwerker alleen gevuld voor flexwerker en einddatum gevuld.
- C:Domain constraint RDNEINFLEX violated, invalid value 3

Violations:
- C:Constraint HIS_IKV_IKGH02 violated: Er mag geen overlap tussen de bestaanstijden in de historie zijn.

4. De overheid: autonomie versus integratie

Nationaal beleid

1. Basisregistraties (BRS): belangrijke gegevens op 1 plek, steeds beheerd door 1 partij. Eenmalige uitvraag. Gebruik verplicht. Fraude bemoeilijkt omdat alle overheden dezelfde info bezitten.
2. Gegevensuitwisseling gebaseerd op globale sleutels (bijv. BSN).
3. Fouten en verdachte data terugkoppelen.

Gangbaar lokaal gedrag

1. Liefst zelf data verzamelen, BRS data sowieso lokaal opslaan. Ketens zijn *procesketens* waarin de data van database naar database “hoppen”.
2. Weinig of niet terugkoppelen naar BRS houder.
3. Lokale databases met beperkte scope vrijwaren van spiedende blikken van derden, *juist als de gegevenskwaliteit slecht is*.

Status 2012:

Publieke organisaties aan winnende hand. BRS projecten mislukken. Gegevenslogistiek loopt niet cf. BRS stelsel. Wel deelsuccessen sleutelharmonisatie (wel BSN maar bijv. geen uniforme identificatie van bedrijven). *Corruptie in enge en ruime zin is endemisch.*

5. De Amsterdamsche spookburger casus: ommakeer?

1. UWV bezit alle adressen van economisch actieve personen via de loonaangifte. De kwaliteit daarvan is super (beter dan GBA).
2. Voor alle 14 miljoen personen die UWV registreert houdt zij (volstrekt overbodig) een kopie uit de GBA bij.
3. In een onbewaakt moment worden de bestanden op verzoek van A'dam vergeleken en blijken 131.000 [!] mensen voor wie aangifte wordt gedaan niet bekend in de GBA danwel elders woonachtig.
4. A'dam wil per se haar GBA op orde krijgen (atypisch) en vraagt UWV om de gegevens. UWV en SZW weigeren.
5. RTL Nieuws bericht over de verschillen en laat zien dat de grootste bron van de fouten is UWV zelf!
6. UWV (nieuwe RvB) heeft nu aandacht voor de datakwaliteit problemen, zeker nadat blijkt dat er “slechts” 64.000 adresverschillen zijn.
7. *Het Kabinet gaat de wetgeving aanpassen zodat publieke organisaties zich niet meer gemakkelijk andere organisaties die hun gegevens willen valideren buiten de deur kunnen houden. Zou het helpen?*

6. Naar een integere publieke gegevenshuishouding

Kenmerken huidige situatie publieke sector:

1. Versnipperde lokale databases met soms globale sleutels.
2. Lokale databases zijn in enge zin vaak redelijk integer maar in ruime zin vergaand corrupt (fout en/of onvolledig). Soms is fraude daarvan de oorzaak en soms het gevolg. Hoe dan ook neemt de overheid suboptimale beleidsbeslissingen.
3. Globale sleutels maken verzamelingen van lokale databases tot één samenhangende (logische) globale database. *Tekortschietende integriteit in ruime zin wordt daardoor opeens tekortschietende integriteit in enge zin.*
4. Techniek van bestandsvergelijking is nog nooit zo eenvoudig geweest.
5. De wet wordt aangepast.

Gaan overheidorganisaties nu hun database content onderling vergelijken en verbeteren? Onwaarschijnlijk, tenzij daartoe worden gedwongen, liefst natuurlijk met zachte hand. Men is bang voor negatieve publiciteit en heeft vaak weinig belang bij verbetering en fraudebestrijding.

6. Voorstel voor een zachte hand: de Trias Informatica

Snelle verbetering vraagt om keiharde dwang van boven of om zachte tegenkrachten. Het laatste heeft de voorkeur.

Voorstel voor een Trias Informatica (globale versie):

1. Maak jaarlijks in de tijd een afslag van alle belangrijke publieke databases en verzamel deze in een analyseomgeving: GBA, NHR, BVR, Polisadministratie, NTS, Landbouwsubsidiesysteem, ziektekostensystemen, etc.
2. Vergelijk de data en koppel geconstateerde verschillen ter verklaring/opheffing terug naar de betreffende organisaties, *naar een zwaar bemenste stuurgroep onder AZ. Rapporteer in 2^e instantie ook naar de Kamer.*

Voorspelde gevolgen:

1. Organisaties gaan energiek hun databases op orde krijgen.
2. Organisaties gaan hun processen evalueren en verbeteren, niet in de laatste plaats door gebruik te maken van basisgegevens omdat ze dan minder kwetsbaar zijn voor kritiek.
3. Miljardenmeevallers op de begrotingen: doordat fraudes boven water komen, burgers en bedrijven hun gedrag heroverwegen en de overheid efficiënter wordt.

Dank voor uw aandacht

rene.veldwijk@ockham.nl
www.ockham.nl