Belastingdienst

Referentiearchitectuur Gegevens

Editie: Q1 2019

Uitgavedatum: 12 april 2019

Status: Concept

Inhoudsopgave

Metagegevens

Opdrachtgever

Naam	Organisatie	Rol
Maarten Jonker	Belastingdienst IV&D	Chief Data Officer

Auteurs

Naam	Organisatie	Rol
Ronald Damhof	Belastingdienst IV&D	Enterprise Data Architect

Versiebeheer

Versie	Datum	Status	Wijzigingen	Auteur
Q1 2019	19 maart 2019	Concept	Opleverversie	Ronald Damhof
Q1 2019	12 april 2019	Concept	Reviews M.Jonker, M.de Bruijne, M.Koster, A.Karels, EA review (hoofdstuk 6 toegevoegd, hoofdstuk 5 uitgebreid)	Ronald Damhof

Figuren:

1. Achtergrond

Er zijn twee wezenlijke ontwikkelingen gaande, zowel in het maatschappelijke debat als in de Belastingdienst bestuurlijke context, waar gegevens een cruciale rol spelen in het realiseren van ambities die zijn gesteld.

Burger & Bedrijf in regie op haar gegevens

Zowel in het publieke als private domein is m.b.t. persoonlijke gegevens een discussie op gang gekomen die steeds meer materialisatie krijgt. Dit is zichtbaar in het maatschappelijke, politieke en ook juridische debat. De kern van die discussie is dat subjecten[1] regie op hun eigen gegevens moeten gaan krijgen. In het publieke domein zien we dat enerzijds door ontwikkelingen als 'Regie op Gegevens[2]', 'Digikluis[3]' en 'Gebruiker Centraal[4]' en anderzijds door wetgeving zoals de AVG[5] en de WOO[6]. Ook in het semipublieke en private domein zien we in verschillende sectoren soortgelijke discussies[7].

In de context van de Belastingdienst is dezelfde discussie relevant in het kader van het minimaliseren van de tax gap[8]. Het gaat dan met name om het maximaliseren van de compliantie. De strategie rept daar dan ook expliciet over[9]:

> *We streven er naar dat burgers en bedrijven bereid zijn uit zichzelf (fiscale) regels na te leven, zonder dwingende en kostbare acties van de Belastingdienst.*

Ook in de beleidsvisie Integraal datamanagement[10] is expliciet het volgende geformuleerd in principe #4:

> *De Belastingdienst is transparant over en aanspreekbaar op de gegevens die hij verwerkt. Hij streeft ernaar burgers en bedrijven zo veel mogelijk actief inzicht te geven in hun gegevens en hen regie te geven op eigen gegevens. Gegevens worden, binnen wettelijke grenzen, als open data beschikbaar gesteld.*

---

[1] Met subjecten worden natuurlijke, niet-natuurlijke en rechtspersonen bedoeld dan wel (in juridische zin) de vertegenwoordigingsbevoegde

[2] https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/regie-op-gegevens/

[3] https://www.digitaleoverheid.nl/achtergrondartikelen/de-persoonsdatakluis-en-regie-op-eigen-gegevens/

[4] https://www.gebruikercentraal.nl/

[5] Algemene Verordening gegevensbescherming https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/avg-europese-privacywetgeving/algemene-informatie-avg, dataminimalisatie, art.15 inzagerecht, etc..

[6] Wet Open Overheid, https://www.eerstekamer.nl/wetsvoorstel/33328_initiatiefvoorstel_snels_en

[7] Zorg: PGO (persoonlijke gezondheidsomgeving), medmij.nl, Banken: PSD2

[8] "The difference between total amounts of taxes owed to the government versus the amount they actually receive. Generally, a tax gap is caused by taxpayers overstating deductions and understating their income so they can pay fewer taxes; but late paying taxpayers also cause the tax gap." Read more: http://www.businessdictionary.com/definition/tax-gap.html

[9] Jaarplan 2019, Belastingdienst, https://www.rijksoverheid.nl/documenten/jaarplannen/2018/11/05/jaarplan-2019-belastingdienst

[10] Beleidsvisie Integraal datamanagement, 20-11-2017 door DT BD goedgekeurd

Vertaald naar gegevens betekent dat bijvoorbeeld dat het subject op continue basis in staat moet zijn om de voor haar relevante gegevens te kunnen inzien, valideren en wijzigen. Waarbij de fiscale consequenties transparant en begrijpelijk zijn.

Sterker nog, subjecten moeten zelf in staat zijn om te kunnen zien met wie de Belastingdienst haar gegevens heeft gedeeld, wanneer en onder welke grondslag. Nog een tandje erbij; zou het subject niet zelf in staat moeten zijn om bijvoorbeeld haar fiscaal inkomen tijdelijk te delen met de hypotheekverstrekker of de verhuurder?

Anders geformuleerd; gegevens in zijn algemeenheid, maar met name gegevens rondom de context van het subject zijn een cruciale capability die zowel maatschappelijk als bestuurlijk (t.a.v. de strategie van de Belastingdienst) van groot belang is.

Gegevens zijn dus de spin in het web van deze ontwikkelingen

Hoe triviaal dit statement ook klinkt, het is een hele spannende omdat van oudsher de middelenadministraties ingericht zijn naar de belastingwetten. Wat hier wordt gevraagd is dat te eerbiedigen[11] maar het gegevensdeel van de middelenadministraties meer en meer te kantelen naar de belangen van het subject; de burger en het bedrijf.

Uitnutten van technologie

Er is een technologische acceleratie gaande op het gebied van artifical intelligence, machine learning, natural language processing, etc.. Datascience, het samenvattende vakgebied, kan de Belastingdienst grote voordelen bieden, zowel in efficiëntie van processen, maar zeker ook t.a.v. een verhoging van de kwaliteit van dienstverlening en bij het uitoefenen van haar primaire taken.

De broedkamer, later DF&A, en ook aansprekende voorbeelden bij Toeslagen, Douane, FIOD en IV-accent hebben dit herhaaldelijk aangetoond. De neiging om te investeren in de mensen en middelen aangaande deze ontwikkelingen is dan ook begrijpelijk. Echter, er zijn twee uitdagingen die een risico vormen t.a.v. het halen van voldoende rendement uit datascience:

1. De kwaliteit van de belangrijkste grondstof wordt vaak over het hoofd gezien: data.

   'Dat regelen we wel', 'dat programmeren we wel aan elkaar'…..

   Dit is een schromelijke onderschatting van de inherente complexiteit van data die altijd in een context leeft, altijd bevooroordeeld is en bovendien van inferieure kwaliteit t.o.v. van de daadwerkelijke leefwereld van burger en bedrijf.

[11] Idealiter zou je uitvoering en wetgeving meer als spreekwoordelijke 'brothers-in-arms' willen zien optreden, zie ook Mariette Lokin; wendbare wetgeving EAN: 9789462905528

Het gevolg is dat elke medewerker met een datascience profiel, ieder op zijn eigen manier, zich voor het grootste deel van zijn of haar tijd bezig moet houden met het opwerken, schonen, valideren, integreren en transformeren van data. Voordat er ook maar één algoritme is gemaakt[12].

Dat zou niet zo moeten zijn, werkt demotiverend en is risicovol voor de Belastingdienst aangaande het behouden van deze zeer schaarse competenties.

2. Gelieerd aan bovenstaande is het grote probleem in datascience waar publiek en private organisaties allemaal mee kampen: de operationalisatie ervan t.b.v. het primaire proces en/of taak. Te vaak blijven briljante algoritmes in de 'garage' omdat niet van te voren is nagedacht wat het betekent om bv. een risicomodel op schaal- en beheersbaar in productie te nemen. Alle aspecten van software engineering komen daarbij kijken, denk aan informatiebeveiliging, interoperabiliteit, exploitatie, deployment, versionering, verandermanagement, etc..

   Maar denk bij die operationalisatie ook vooral aan de data. Die data moet bij een organisatie als de Belastingdienst onbetwistbaar zijn; gevalideerd, betekenisvol, traceerbaar en auditeerbaar.

   Serieus in datascience investeren en daar voordeel uit willen halen vereist investeringen in haar grondstof; de architectuur, het management en de governance van deze grondstof. En de discipline om dit vol te houden.

[12] Expert meningen van Michael Stonebreaker, Andy Bitterer en ook onderzoeksinstituten als Gartner en Forrester geven aan dat dit in praktijk van 80~90% van de totale datascience capaciteit opslokt en in de weg staat van verdere professionalisering.

2. Leeswijzer

Dit document, geschreven door de concerndirectie IV&D, geeft de kaders t.a.v. gegevens vanuit een organisatie breed perspectief. Het is bovendien een zich evoluerend geheugen dat vastlegt hoe en waarom deze kaders tot stand zijn gekomen. De eerste hoofdstukken vormen een beschrijvende kern en zijn voor alle doelgroepen binnen de Belastingdienst relevant, hoofdstuk 6 geeft een wat meer visuele duiding t.a.v. domein, organisatie en capabilities. Hoofdstuk 7 is relevant voor personen die zich inhoudelijk begeven op het vlak van gegevens.

Basis voor dit document is een presentatie 'Grip op gegevens, de contouren van een visie en strategie'.

De huidige versie van dit stuk heeft, tenzij expliciet anders genoemd[13], een focus op de gestructureerde gegevens, dit zijn de gegevens uit de bronadministraties en/of de gegevens van derden.

T.a.v. de leesbaarheid is het volgende van belang:

[1.] 'Gegevens' en 'data' worden in dit stuk door elkaar gebruikt en hebben dezelfde betekenis;
[2.] Overal waar 'data' staat, moet ook 'metadata' worden gelezen;
[3.] Daar waar 'Authentieke bron' staat kan ook 'System of Record' worden gelezen en vice versa;
[4.] Het concept 'bronadministratie' is een supertype van het concept 'middelenadministratie'.

Referentiearchitectuur gegevens[14]

De regels/principes in de voorgaande versie van de referentiearchitectuur gegevenshuishouding en informatiepositie worden opgevolgd door de dit document, de gevolgen t.a.v. uitvoering wordt beschreven in de domeinarchitectuur gegevens.

Om zoveel mogelijk transparant te zijn aangaande de status van de regels/principes van de Referentiearchitectuur gegevens, is in Bijlage F een mapping opgenomen tussen de regels/principes uit de vorige versie van de referentiearchitectuur en de huidige versie. Alle regels/principes uit de referentiearchitectuur zijn opgenomen, sommige nader geduid en verzwaard.

Beleidsvisie Integraal datamanagement[10]

Datamanagement en data-architectuur hebben veel afhankelijkheden en raakvlakken. In bijlage G is een overzicht van deze raakvlakken gemaakt van de genoemde datamanagement principes en de kaders van dit document.

[13] In hoofdstuk 5 is een poly-gestructureerde dienst opgenomen die met name over de beheersing van documenten, scans, beelden etc.. gaat.
[14] Referentie Architectuur Ggegevenshuishouding en informatiepositie, december 2016. A. Karels en S.Otten

3. Inleiding

Conform de 'visie op gegevens', in de context van de achtergrond van Hoofdstuk 1, zijn er twee doelstellingen geformuleerd die de komende jaren gerealiseerd moeten worden bij de Belastingdienst:

- Waardecreatie (Data as an asset): Gegevens moeten ervoor zorg dragen dat de Belastingdienst vanuit het perspectief van de het subject[1], ongeacht het middel, gegevens ter beschikking kan stellen t.b.v. burger, bedrijf, medewerker, toezicht, kantoor, sturing en verantwoording, rekening houdende met de quality of services.

- Risicobeheersing (Data as a liability): Gegevens moeten ervoor zorgen dat de Belastingdienst dit op een wijze doet waar risico's t.a.v. gegevens beheerst worden. Denk hierbij aan aspecten als transparantie, datakwaliteit, privacy by design, beveiliging, data minimalisatie, beheersbare verandering, continuïteit en schaal.

Het nastreven van beide doelstellingen gaat hand in hand, ze zijn als het ware de yin en yang van gegevens. De waarde-creërende kant onvoldoende realiseren, wat nu het geval is, leidt tot een steeds grotere toename van de risico's t.a.v. gegevens, met alle gevolgen van dien, bijvoorbeeld:

- De enorme proliferatie[15] van gegevens binnen de Belastingdienst vergt veel middelen t.a.v. continuïteit. Denk aan de vele formele gegevensgerichte systemen zoals Enterprise Data Warehouses, Data Fundamenten, DSS'en en vele lokaal ontwikkelde dataopwerkingen. Maar denk ook de berg aan informele gegevensgerichte systemen[16];
- De enorme proliferatie van gegevens binnen de Belastingdienst weerhoudt haar ervan de gewenste quality of services te leveren;
- De enorme proliferatie van gegevens binnen de Belastingdienst leidt tot vele vormen van subadministraties waardoor burgers en bedrijven 'vast' komen te zitten in het systeem (lees; een opgewerkte, niet geïntegreerde dataset)[17];
- De enorme proliferatie van gegevens binnen de Belastingdienst doet de gebruiker van deze gegevens vragen naar de actualiteit, relevantie, betekenis, accuratesse, oorsprong, etc..;
- De enorme proliferatie van gegevens binnen de Belastingdienst zorgt ervoor dat er moeilijk een up-to-date verwerkingsregister[18] geraadpleegd kan worden die de daadwerkelijk verwerking (in operatie) representeert.

---

[15] Met proliferatie wordt in deze context bedoeld de ongecoördineerde massieve verspreiding van gegevens

[16] ~75TB MS Access databases, ~40TB SAS Datasets, ~20TB FIL/csv bestanden, Excel bestanden en macro's niet meegerekend

[17] Marlies van Eck, geautomatiseerde ketenbesluiten & rechtsbescherming: https://www.recht.nl/nieuws/ict/164318/proefschrift-geautomatiseerde-ketenbesluiten-rechtsbescherming/

[18] https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/20180419_art29wp_position_paper_on_records_of_processing_activities.pdf

4. Root-causes, solutions en aspecten

Met het risico om een zeer complex geheel, overmatig eenvoudig te duiden, is het zinvol om grote lijnen van oorzaak-gevolg te onderkennen.

Root-causes

1. De proliferatie[15] van gestructureerde[19] gegevens is voor het grootste deel een manifestatie van 'dataschuld[20]' in de bronadministraties.

   De belangen van de betreffende applicatie/interactie en het te ondersteunen primaire proces wegen erg zwaar in de wijze waarop gegevens tot stand komen. Gegevens zijn daarmee een bijproduct geworden van de applicatie. Ofwel, de ordening van gegevens in systemen past bij het primaire gebruikersdoel van de administratie. Echter, de belangen van de burger, bedrijf, medewerker, toezicht, kantoor, sturing en verantwoording worden niet of onvoldoende meegenomen[21].

   Vb #1. Het huidige EDW van de Belastingdienst is voor een groot deel een concretisering van dataschuld in de bronadministraties[22]. Veel domeinspecifieke logica is binnen het EDW gerealiseerd (en moet daar dus worden onderhouden) terwijl die logica ook aanwezig is in de betreffende bronadministratie. Echter omdat de output van de logica ofwel omdat de logica zelf niet als service wordt aangeboden, zijn vele partijen binnen de Belastingdienst bezig om deze logica te realiseren.

   Vb #2. De identificatie van subjecten in bronadministraties is zeer gevarieerd[23] en soms zelfs 'overloaded[24]'. Dit geeft enorme integratieproblemen binnen domeinen, laat staan over domeinen heen. De oplossing die vaak wordt opgezocht is het ontsluiten van data (lees; kopiëren) en de integratieproblemen benedenstrooms 'oplossen' met ETL[25] technologie. Afgezien van het risico t.a.v.

[19] Gestructureerde gegevens heeft betrekking op data afkomstig uit onze bronadministraties en/of gegevens van derden (niet zijnde documenten, beelden of id.)

[20] Naar analogie van de definitie van technische schuld: De schuld bestaat uit de kosten voor de herstelwerkzaamheden (Engels: refactoring) die uitgevoerd moeten worden om een consistente en onderhoudbare oplossing te realiseren. Zo lang dit werk niet gedaan is blijft de schuld bestaan en wordt er telkens rente betaald, in de vorm van extra inspanning voor wijzigingen.

[21] Anders geformuleerd; er is onvoldoende tot geen 'separation of concerns' (Edsger Dijkstra) toegepast in het realiseren van de datalaag in bronadministraties. Mariette Lokin noemt dit in haar boek 'Wendbare wertgeving'ook wel een onderscheid maken tussen 'know' en 'flow'

[22] Een vuistregel is vaak; het aantal en de grootte van informele en formele gegevensgerichte (afgeleide) systemen zoals datawarehouse geeft vaak een indicatie van de omvang van de dataschuld in de system of records.

[23] BSN, RSIN, SOFI, FI, etc.. Binnen 1 domein is geconstateerd dat er 27 verschillende vormen (domeintyperingen) van BSN# zijn gebruikt

[24] Overloaded wil zeggen dat een veld in de database wordt overladen met verschillende betekenissen. In het geval van identificaties is geconstateerd dat er databases zijn met een ID veld waar ofwel een BSN, ofwel een RSIN, ofwel een FI ofwel een SOFI in staat. Alleen moeilijk te onderhouden logica bepaald bij elke instantiatie wat het (mogelijk) is.

[25] ETL: Extractie, Transformatie en Laden

extra kopieën en de gebrekkige quality of services (dit soort fysieke integratie kost latency en leidt tot uitval die gemanaged moet worden) is het grootste probleem dat dit probleem niet wordt opgelost, waar het ontstaat (bij de bron) en dus steeds weer leid tot kosten en risico's.

Vb #3. Er zijn bij de Belastingdienst nieuwe systemen gerealiseerd, dat de komende 10 jaar moet draaien. De implementatielaag van data voor enkelen van deze systemen zijn gerealiseerd in een xml opslag of wel een zogenaamd key-value-pair[26]. Vanuit applicatie-belangen gedacht is dit erg wendbaar en flexibel, maar de prijs t.a.v. belastingdienst-belangen (toezicht, sturing en verantwoording, interactie, kantoor, burger, bedrijf, dienstverlening) is hoog. Integratie van deze data over domeinen heen is problematisch alsmede begrijpelijkheid en verklaarbaarheid is lastig tot misschien wel onmogelijk. De root cause is hier niet dat er gekozen is voor opslag technologie of methodiek x of y, maar dat er geen logisch datamodel aan ten grondslag ligt die ge-enforced wordt en een mapping heeft naar het betreffende technische formaat.

2. Daar waar er dataschuld is in bronadministraties, leidt dit tot risico's t.a.v. databeheersing[27]. Hoe groter de dataschuld, des te groter de risico's t.a.v. data. Deze dataschuld resulteert in een explosie van vele Lokaal Ontwikkelde Applicaties (LOA's) die data integreren, combineren en uitleveren op vele verschillende onbeheerste wijzen. Die vaak ook nog in primaire processen worden gebruikt en dus in de continuïteit steeds meer middelen vragen

E.e.a. vat zich samen in drie root causes van de gestelde problematiek:

a) Een slechte ontwerp discipline van informatie- en datamodellen die ook daadwerkelijke de fysieke en gerealiseerde data representeren en alle belangen (concerns) afdekken, met name die van het subject (en als afgeleide dus; interactie met burger & bedrijf, toezicht, dienstverlening en Management Informatie). *Hierdoor is betekenis en herleidbaarheid naar de wet niet mogelijk of zeer intensief en foutgevoelig. Bovendien leidt dit tot onnodige silo-vorming t.a.v. data.*

b) Strakke kaders, processen en methodieken rondom de wijze waarop data-integratie moet plaatsvinden. Zowel binnen bronadministraties als in allerlei opwerkingen. *Hierdoor is een veelheid aan integratie, ontsluitingsmechanismen en kopieën van data ontstaan.*

[26] Simpel geformuleerd zijn dit drie tabellen waar alle data in wordt opgeslagen zonder ooit de database aan te hoeven passen.

[27] Er kan hier een verwijzing worden gemaakt naar de entropische eigenschap van data; zonder beheersing en governance zal data de natuurlijke neiging hebben om zich eindeloos door te kopiëren tot een staat van 'chaos/donker/kou'.

c) Binnen die veelheid van ontsluitingsoplossingen en data kopieën is er geen discipline rondom de positionering van business logica. *Hierdoor worden opwerkingen vaak onterecht onderdeel van een primair ketenproces.*

Root-solutions

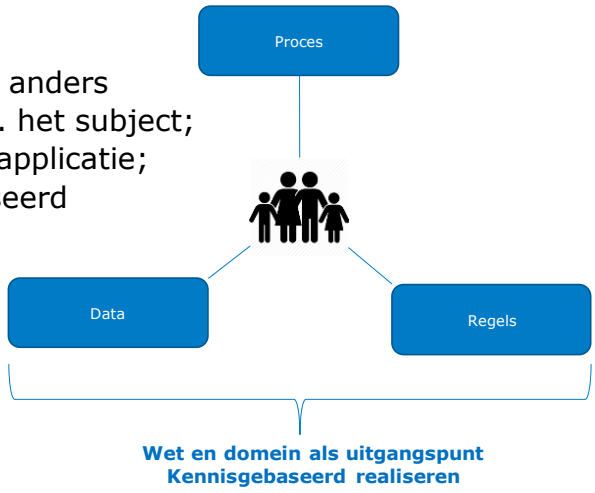
Met het risico om een 'wicked[28] problem' af te schilderen als 'eenvoudig' en dat er dus een aantal 'silver bullets' zouden zijn, kan de kern van een oplossingsrichting zich als volgt samenvatten:

1. Data wordt opgehaald bij de system of record en ook daar gewijzigd[29];
2. Dataschuld wordt bij de authentieke bron opgelost, niet benedenstrooms;
3. Rekenregels (logica) gebruiken data die de system of record niet verlaten heeft;
4. Alle vormen van dataleveringen[30] (data in beweging) staan onder governance[31];
5. Alle vormen van dataopslag (data in rust) staan onder governance[24;]
   Een compromisloze strategie t.a.v. standaardisering van informatie- en datamodellen en de instantiaties van data.
6. Stop met het kopiëren van data tenzij daar technische concerns voor gelden.

Anders geformuleerd

1. De manier waarop data wordt opgeslagen moet anders
   - Rekening houdende met alle belangen , m.n. het subject;
   - De wet/domein is het uitgangspunt, niet de applicatie;
   - "Know & Flow" uit elkaar halen, kennisgebaseerd werken[32];
   - Data naar analogie van regelbeheer[33];
   - Data in rust onder forse governance.

2. Stop met het kopiëren van data als daar geen technische reden voor is
   - Van ontsluiten naar aansluiten;
   - Alle bewegingen van data onder governance.



[28]Wikipedia: A wicked problem is a problem that is difficult or impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize.

[29] Aansluiten ipv ontsluiten; dit principe is bepaald niet nieuw en kent haar oorsprong in de basisadministraties en ook nieuwere bewegingen zoals 'Common Ground' van de VNG (https://vng.nl/samen-organiseren/common-ground)

[30] Een API op bron X die wordt gebruikt door bron Y is ook een vorm van gegevenslevering

[31] Het woord 'governance' moet hier worden gelezen in ruime vorm, ofwel alle maatregelen die nodig zijn om het leidend voorwerp bestuurbaar te maken

[32] 'De kennis die nodig is voor de uitvoering van de taken wordt gescheiden van het systeem waarmee deze geëxecuteerd wordt' – Wendbaar wetgeven, M.H.A.F. Lokin, 2018.

[33] Verwezen wordt hier naar de Business Rule Management werkzaamheden aangaande de implementatie van logica die wordt afgeleid van wetteksten.

Root-aspecten

De volgende aspecten t.a.v. gegevens worden beschouwd als differentiërend en zijn integraal onderdeel van de kaders op gegevens zoals beschreven in hoofdstuk 8:

1. Modelleren	Alle *state* van data zijn gemodelleerd. De modellen zijn geversioneerd, vindbaar, doorzoekbaar en representeren *altijd* de daadwerkelijke situatie in productie.
Modellen zijn metadata	Modelinformatie is cruciale metadata.
Natuurlijke taal	Alle informatie- en of datamodellen zijn altijd terug te herleiden tot natuurlijke taal die domein experts, in combinatie met concrete voorbeelden, kunnen valideren[34].
Data is leidend, niet de applicatie	We modelleren domeinen (universes of discource)[35], geen applicaties. Data is geen bijproduct van een applicatie, sterker nog, precies andersom. Domeinen worden gemodelleerd en daarna de interactie/logica.
Precies modelleren en operationaliseren	Er worden geen informatie- en/of datamodellen ontworpen die niet geoperationaliseerd worden. Ofwel; formele informatie- en of datamodellen zijn zeer precies en omkaderd door een methodologie (publiekelijk beschikbaar) een notatie (incl. gereedschappen) en een aanpak.
CGM's zijn schetsen	Conceptuele Gegevens Modellen (CGM's) die 'ongeveer' geïmplementeerd zijn, zijn niet toegestaan als een formeel ontwerp-artefact, hoogstens als een *schets* om discussies te voeren.
Nieuwe context/realiteit wordt gemodelleerd	Integratie van data over domeinen (modellen) heen resulteert *altijd* in een nieuwe context/realiteit. Nieuwe context moet altijd expliciet gemodelleerd worden.

[34] Dit principe wordt reeds toegepast bij Business Rule Management

[35] Een Universe of Discourse kan een wetgeving zijn, een te ondersteunen bedrijfsproces, etc..

2.	Multirealiteit	*Eén versie van de waarheid is een gevleugelde uitspraak in organisaties.* *En is grote onzin.* Een doorsnee grote organisatie is te complex voor één versie van de waarheid. Er circuleren altijd meerdere versies van de waarheid. De waarheid wordt altijd contextueel bepaald. Dat worden ook wel *multirealiteiten* genoemd. Het is in de moderne data-architectuur de heilige graal. Alle feiten behoren tot een bepaalde waarheid, sommige feiten behoren tot meerdere waarheden. Je hebt dus meerdere realiteiten voor dezelfde feiten. Multirealiteiten maken wel gebruik van dezelfde feiten. Er moeten twee typen mutirealiteit onderscheiden worden: ▪ Horizontaal: binnen een universe of discourse/domein van gegevens/kennisdomein (M.Lokin) maak je met verschillende afleidingsregels andere contexten/realiteiten. Dit is wat Lokin noemt; scheidt 'know' en 'flow', dit wil je met name ondersteunen in het bronnen-landschap. ▪ Verticaal: feiten uit verschillende universes of discourse/domeinen van gegeven/kennisdomeinen combineer tot een volstrekt nieuwe context/realiteit. Iets wat veel gebeurd binnen BI en anaytics (set-gebaseerde dienst, zie hoofdstuk 5).
3.	Tijdreizen en Reizende-nu	Er is een onafhankelijke tijdsconsistente toegang tot data/informatie/metadata mogelijk, zowel transactioneel als batch. Dit wil zeggen: onafhankelijk van wanneer je data op een tijdstip in het verleden opvraagt, krijg je altijd dezelfde data (los van het feit dat de data op continue basis een andere actualiteit kent). Dit laatste wordt ook wel omschreven als *reizende-nu.* Anders geformuleerd; Volledige temporele non-destructieve verwerking van data en metadata.
4.	Privacy by Design & Default[50]	*a) Dataminimalisatie*

Data minimalisatie is het beperken van data gebruik tot alleen die gegevens die noodzakelijk zijn om het beoogde doel te bereiken.

Een van de zeven beginselen uit de AVG (artikel 5 lid 1 c) betreft dataminimalisatie: Persoonsgegevens moeten toereikend zijn, ter zake dienend en beperkt tot wat noodzakelijk is voor de doeleinden waarvoor zij worden verwerkt („minimale gegevensverwerking").

Heel concreet betekent dit dat er alleen gegevens worden gevraagd die dekkend zijn voor het gestelde en geformuleerde doel en waar bovendien een wet of convenant aan ten grondslag ligt.

Daarnaast worden gegevens niet langer bewaard dan nodig.

Dataminimalisatie is de nieuwe standaard geworden in de algemene omgang met gegevensverzamelingen. Ook het intern vragen naar of ter beschikking stellen van data intern valt onder het dataminimalisatie principe. Een relevante vraag in dit verband is; wat heb je precies nodig voor je werk? Niet: 'geef mij alles maar'.

b) Pseudonimisering
Het veilig en verantwoord omgaan met informatie is een zaak waar we ons allemaal voor inzetten. Pseudonimiseren is daarbij één van de technische beheersmaatregelen die bijdraagt aan het beschermen van gevoelige gegevens. Toegang tot gegevens wordt al beperkt door het gebruik van autorisatie-profielen en data minimalisatie. Deze profielen verhinderen ongewenste inzage door onbevoegden en verschaffen medewerkers toegang tot slechts die gegevens die noodzakelijk zijn voor de uitvoering van de dagelijkse werkzaamheden. Waarom dan toch ook pseudonimiseren?

- Door middel van pseudonimiseren wordt het direct identificerende karakter van gegevens verborgen. Hiermee wordt de privacy van de betrokkene(n) beter beschermd ten tijde van de verwerking;

- Pseudonimiseren met behulp van een code-boek biedt een extra bescherming bij verlies en/of

	diefstal van gegevens omdat het code-boek die nodig is om het pseudonimiseren ongedaan te maken uniek zijn voor de organisatie en niet bij de gegevens worden opgeslagen; ▪ Pseudonimiseren is een maatregel die bijdraagt aan het voldoen aan wet- en regelgeving met betrekking tot de verwerking van persoonsgegevens (artikel 25 en 89); Pseudonimiseren is niet volledig waterdicht, vaak kan, in combinatie met andere informatie toch herleid worden naar een individu. Hiertoe heeft de directie vaktechniek een nota[36] geschreven t.b.v. Privacy by Design bij DF&A, deze nota is ook onderdeel van de wijze van pseudonimiseren in dit document/. *c) Vastleggen AVG aspecten* Bij alle state van data moet worden vastgelegd wat de dataclassificatie, grondslag, gegevensverantwoordelijke, doelbinding en 'te pseudonimiseren[37]' attributen zijn.
5. Transparantie en Auditeerbaarheid	110% Transparantie is dat ***alles*** binnen het operationele proces qua dataverwerking en uitvraag op een uniforme manier inzichtelijk gemaakt kan worden, en dat we dit kunnen controleren (auditeren) om zo zelfs problemen binnen het proces zelf te zien, niet alleen de verwerkte data. Bovendien stelt het ons in staat om ten alle tijden de beschikking te hebben over een up-to-date verwerkingsregister[38]. Dit aspect moet ondersteunt worden door operationele aspecten zoals beschreven in bijlage I.
6. Datakwaliteit by Design	Informatie- en datamodellen moeten operationaliseerbaar zijn. Door informatie- en datamodellen formele ontwerp-artefacten te laten zijn wordt met name de definitie, integriteit, onderlinge

[36] Directie vaktechniek, memo pseudonimiseren, 25 maart 2018

[37] Privacy by Design, DF&A, 4-12-2018, versie 1.01:Fiscaal nummer (en velden waarvan dit nummer is af te leiden, zoals BSN), Voornaam, Tussenvoegsel(s), Naam (achternaam of bedrijfsnaam), Straat, Huisnummer, Huisnummer toevoeging, Postcode, plaats

[38] Een verwerkingsregister vanuit een AVG perspectief heeft alleen betrekking op gegevensverwerkingen van persoonsgerelateerde gegevens. Verwerkingsregister in de hier beschreven context gaat over alle gegevensverwerkingen (waar gefilterd kan worden op gegevensverwerkingen van persoonsgebonden data)

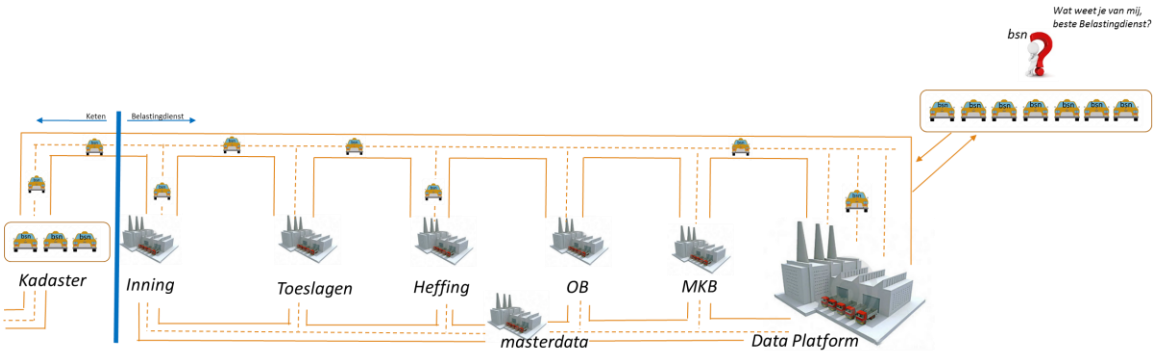
	consistentie en accuratesse van data (cruciale aspecten t.a.v. datakwaliteit) meer gegarandeerd. Als er sprake is van een gegevenslevering waar de Belastingdienst zelf niet de authentieke bron is, dan wordt er *altijd* gevalideerd op minimaal een logisch datamodel. Er vindt altijd feedback plaats van de validatieresultaten naar de aanleveraar en er is van te voren afgesproken welke type validaties direct door de aanleveraar opgelost moeten worden (bv. blokkerende validaties) en welke t.z.t. opgelost moeten worden (bv. signalerende validaties). Blokkerende validaties hebben altijd een procesconsequentie; levering niet geaccepteerd, aka 'ketenpartner, u heeft niet aan uw verplichting voldaan'. Dit moet vooral worden gezien als een verdere professionalisering van data-uitwisseling door elkaar aan te spreken op rechten en plichten. Belangrijke nuance in dit geheel is dat de balans tussen de *valideren en blokkeren* versus *valideren en beschikbaarstellen* door de wet kan zijn bepaald. Bijvoorbeeld, de Belastingdienst MOET alle aangiften accepteren. Alleen op deze manier wordt er een graduele verbetering gemaakt van de datakwaliteit t.a.v. gegevens (bv. van derden) niet alleen binnen de Belastingdienst, maar ook in de keten.
7. Verantwoordelijkheid i.p.v. eigenaarschap	Binnen de Belastingdienst moet voorkomen worden dat *gegevenseigenaar* een term gaat worden. Eigenaarschap van gegevens is een moeizame discussie en juridisch lastig. Het gaat hier om de verantwoordelijke binnen de Belastingdienst, wie is aanspreekbaar op de gegevens, bv. de betekenis/definitie, kwaliteit, classificatie en de verwerking van gegevens. Niet alleen verantwoordelijk voor de gegevens zoals ooit geregistreerd of ontvangen dus, maar ook verantwoordelijk voor de wijze waarop de data verder wordt verwerkt en gebruikt.

5. Te realiseren gegevensdiensten (capabilities)

Er wordt een onderscheid gemaakt tussen drie typen gegevensdiensten die bij de Belastingdienst gerealiseerd dienen te worden.

Subject-gebaseerde gegevensdienst

Als een actuele informatiepositie van een subject[39] vereist is, kan deze per direct worden opgevraagd. Deze dienst wordt minimaal[40] geïnitieerd door middel van een identificerende sleutel (bv. BSN#) en kenmerkt zich door het geven van een resultaat die altijd de huidige stand van zaken weergeeft binnen de Belastingdienst en daarbuiten. Zoals onderstaand figuur aangeeft is het uitgangspunt dat system of records van de ketenpartners onderdeel zijn van de service.



Figuur 1: subject-gebaseerde gegevensdienst naar analogie van een taxi-service

Figuur 1 geeft impliciet al een aantal uitgangspunten weer:

- De gebruiker gaat naar de data toe in plaats van data naar de gebruiker;
- Domeinen/ketenprocessen hebben m.b.t. de data en logica, die behoren bij dat domein/ketenproces, een eigen verantwoordelijkheid om die te leveren;
- Omdat de quality of services hoog moet zijn (near realtime) MOET de data altijd opgehaald worden bij de bron waar de gegevens zijn ontstaan;
- De wijze waarop data wordt opgeslagen (state) en de wijze waarop data beweegt (flow) MOET gestandaardiseerd zijn;
- De wijze waarop verwezen wordt naar domeinoverstijgende concepten (bv. subject) moet geharmoniseerd worden;
- Standaardisatie vereist institutionalisatie[41] en governance van data;
- Alle concepten en instantiaties van concepten die middelenoverstijgend zijn (zoals de identificatie van subject) moeten centraal gemanaged en gegoverned

[39] De dienst heeft haar naam gekregen omdat het subject bij verre de belangrijkste sleutel is waar op gezocht moet worden. Vanzelfsprekend kan de dienst ook worden ingezet voor identificaties van andere objecten, denk aan een kenteken van een auto een zaak# van een dossier.

[40] Er kan sprake zijn van meerdere variabelen die worden meegenomen in het verzoek om data

[41] Binnen de belastingdienst is in de topstructuur op strategisch niveau een CDO aangesteld, op tactisch en operationeel niveau is de 'datafunctie' verspreid

worden; ook wel de *master- en referentie data gegevensdienst* genoemd. Alle middelen hebben een verplichte winkelnering bij deze gegevensdienst.

Set-gebaseerde gegevensdienst

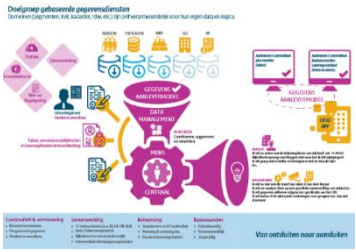
Er zijn binnen de Belastingdienst bedrijfsfuncties en processen die domeinoverstijgende data-integratie vragen en afhankelijk van hun vraag een 'realiteit' – weerspiegeld door data – tot zich willen nemen. Een voorbeeld is een management informatie rapportage waarbij de data een integratie eis over domein heen inclusief specifieke logica om een Key Performance Indicator uit te rekenen. Een ander voorbeeld is dat toezicht een fraudeprofiel wil maken van zzp'ers met een auto op de zaak. Denk ook aan het ontwerpen en trainen van een risicomodel die gebruikt moet worden in de Risico Detectie Service[42] t.b.v. inkomensheffing en waar veel contra-informatie gebruikt wordt.

Kernpunt van deze dienst is dat het een geïntegreerde[43] feitenbasis kent

Alle vormen van analyse[44] op groepen (sets) van gegevens vereisen hoogstens een technische opwerking van gegevens met uitzondering van domeinoverstijgende logica. Deze set-gebaseerde gegevensdienst kent twee subtyperingen:

a) Een gestructureerde variant

Dit is de zogenaamde *Belastingdienst Data Fabriek*[45] (BDF) die in staat is om data uit bronadministraties en data van ketenpartners op zeer systematische wijze voort te brengen en te gebruiken voor set-gebaseerde analyses. De 'produktielijnen' van deze fabriek zijn:



Figuur 2: Een grafische representatie van de Belastingdienst Data Fabriek

- Ontvangststraat; gegevens ontvangen, technisch prepareren voor verdere verwerking en pseudonimiseren;
- Validatie en integratiestraat; gegevens valideren (incl. feedback) en integreren;
- Lever- en ter beschikkingstellingsstraat: gegevens extern leveren[46] en/of ter beschikking stellen alsmede C(R)UD[47] services voor het doen van correcties.

---

[42] Dit is een service welke aangeboden wordt door het domein analytics

[43] Daarmee bedoelen we dat er een feit repository is waar de data semantisch equivalent en temporaal staat opgeslagen en waar de domeinoverstijgende sleutelintegratie is uitgevoerd. Vanuit deze feit-repository kunnen meerdere uitlevermodellen (multirealiteiten) worden gerealiseerd tbv vele verschillende gebruikerstyperingen (analytics, MI/BI, etc..).

[44] Dit is een zeer ruimte interpretatie van 'analyse'; voorspellend, voorschrijvend en beschrijvend vallen er allemaal onder. Ook de 'query 'praktijk op bronadministraties valt hieronder.

[45] Deze variant van de set-gebaseerde gegevensdienst overlapt met het IVG proces, het EDW, deels met de datafundamenten en alle andere vormen van domeinoverstijgende gegevensintegratie.

[46] De voorkeur is om gegevensleveringen naar derden niet meer fysiek te leveren, maar op te laten halen, wanneer nodig door derden (ter beschikking stellen), dit conform Integraal datamanagement principe #5

[47] Create, Read, Update, Delete op een wijze die temporaal non-destructief blijft (en dus traceerbaar en auditeerbaar)

Alle vormen van leveren en ter beschikkingstellingen zijn geadministreerd en worden gemaakt o.b.v. zogenaamde gegevensleveringsovereenkomsten (GLO). Hierin worden de verwachtingen en verplichtingen van de ontvanger ge-expliciteerd. Dit betreft zowel de technische uitwisselprotocollen als syntax als semantische gegevensdefinities.

De fabriek stelt data ter beschikking volgens de hoogste quality of services aangaande privacy by design & default, informatiebeveiliging, tijdreizen, traceerbaarheid, datakwaliteit bij design, etc..

Randvoorwaarden:
- De BDF is DE system of reference voor gegevens van derden[48];
- Alle system of records zijn verplicht een replica van hun gegevens aan deze fabriek ter beschikking te stellen;
- De BDF serviced alleen domeinoverstijgende integratie en logica[49].

Thema: *Pseudonimisering*
Binnen DF&A heeft het datateam de afgelopen jaren veel ervaring opgedaan met het realiseren van een cruciale *Privacy Enhancing Technique* voor de Belastingdienst: pseudonimisering[50]. De set-gebaseerde gegevensdienst moet deze ervaringen meenemen en bevindingen in dit kader overnemen en ter beschikking stellen aan de gehele Belastingdienst.

Thema*: Stop ongecoördineerde query praktijk*
Het ongecoördineerde beslag van allerlei queries en uitvragingen rechtstreeks op de bron moet deze dienst tegengaan. De ontvangststraat (zie hoofdstuk 8, architectuurkaders, Data Logistieke Architectuur) bestaat uit een bronlaag (type 2 – zie kader [DLA.1]) en een gegevenslaag. Deze gegevenslaag is ook de laag waar gegevens *gepseudonimiseerd* worden. Dit is de primaire laag waar opvragingen uitgevoerd kunnen worden.

Alle lagen vóór de gegevenslaag worden afgesloten. In de zogenaamde bronlaag mag alleen data worden opgevraagd welke middels een beheerd artefact (api, scherm, ontsluiting) wordt aangeboden.

[48] De authentieke bron van gegevens van derden (banken, verzakeraars, etc..) ligt bij de betreffende entiteit zelf. In een ideaal eindscenario hoeft/wil de Belastingdienst dit soort gegevens niet meer naar binnen trekken, maar ophalen wanneer nodig bij de betreffende entiteit. Alhoewel dit een prima streven is en er entiteiten zijn die dit ook al kunnen, is de verwachting niet dat dit de komende jaren volledig te realiseren is. Uitvraging (in batch) van gegevens blijft een noodzakelijke dienst.
[49] De BDF levert dus geen system to system integratie als het gaat om gegevens van domein X naar domein Y. Voor de duidelijkheid: als het gaat om een integratie van data over domeinen heen naar een ander domein, dan speelt de BDF wel een rol.
[50] Privacy by Design, DF&A, 4-12-2018, versie 1.01

Thema: *Opportunistische datavoortbrenging mag, onder voorwaarden*
De set-gebaseerde gegevensdienst kenmerkt zich door een systematische, beheerste, robuuste en onder (data) architectuur gerealiseerde opwerking van gegevens die voldoet aan hoge eisen t.a.v. traceerbaarheid, multirealiteit en tijdreizen.

Het is echter ook een voortbrenging die soms niet fit-to-purpose is voor het gevraagde informatieprodukt. Bijvoorbeeld een eenmalige analyse van de panamapapers of een beleidsbeïnvloedend onderzoek naar de schuldenposities van alleenstaanden met kinderen. Denk ook aan het trainen van innovatief voorspelmodel t.a.v. inningen waarbij ook allerlei contra-informatie wordt gebruikt. Dit soort informatieproducten vragen een andere modus operandi aangaande het leveren van gegevens, opportunistische, wendbaarder, kort-cyclischer. Met name bij vraagstukken in het kader van analytics is dit gewenst.

Buiten de BDF[51] mag data worden opgewerkt en geleverd/ter beschikking worden gesteld onder strikte voorwaarden[52]:

- Input is minimaal de gegevenslaag van de BDF;
- Verdere opwerking mag alleen buiten de BDF als het te maken informatieproduct[53] experimenteel (bv.analytics), eenmalig (bv. Panamapapers) ofwel een korte deadline (bv. tweede kamer vragen) kent;
- Alle eisen t.a.v. privacy en informatiebeveiliging blijven onverkort gelden;
- Het zijn gegevensleveringen en deze dienen dus onderdeel te zijn van het verwerkingsregister;
- Resulterende informatieproducten hebben een van de volgende statussen; wanneer wordt het verwijderd/gearchiveerd dan wel wanneer gaat de BDF het realiseren en leveren;
- De verantwoordelijkheid voor de realisatie (en de gevolgen van gebruik) van deze informatieproducten liggen altijd bij de maker (nooit bij de BDF)
- De gebruiker van deze informatieproducten mag de data onder geen beding verder leveren;
- De belangrijkste randvoorwaarde is dat alle opportunistische opwerking van data *gecompartimenteerd* is:

  > *Een besloten ruimte voor zowel de programmatuur als de data behorende bij een functioneel geheel*[50]

Hiermee wordt een invulling gegevens aan twee cruciale non-functionals in gegevens; systematische en robuuste voortbrenging, rekening houdende met de hoogst mogelijk quality of services en lage risicobereidheid versus opportunistische

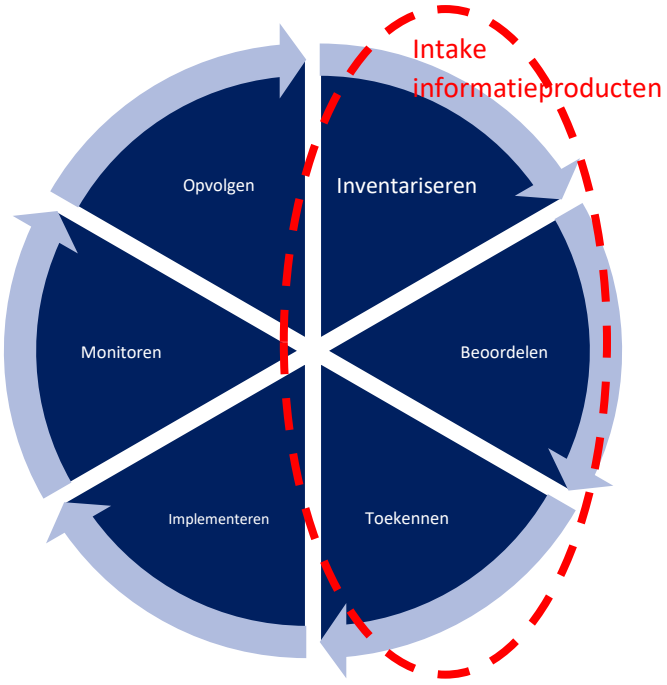
[51] Bijvoorbeeld t.b.v. analytics of directe queries/opvragingen uit bronsystemen
[52] Dit wordt ook wel opportunistische voortbrenging van gegevens genoemd
[53] Een informatieproduct is een combinatie van data en functionaliteit en kan (bijvoorbeeld) variëren van een management informatie rapportage, een risicomodel of een dataset die aan een derde geleverd wordt

en wendbare voortbrenging van gegevens tegen lagere quality of services met een beperkter inzetgebied en hogere risicobereidheid.

Thema: *Vraaggestuurde en risico-gebaseerde informatieproducten t.a.v. de domeinen analytics en gegevens*

Informatieproducten bestaat uit een combinatie van data en een toepassing. Dit kan een management informatie rapportage zijn, dashboard, analytisch model, gegevens t.b.v. analyses, gegevenslevering of een vorm van interactie.

De domeinen gegevens en analytics leveren per definitie altijd een informatieproduct. De wijze waarop dit product tot stand wordt gebracht – a.k.a. het voortbrengingspatroon, is cruciaal om zowel doelmatig als effectief te zijn.



Het bepalen van dit voortbrengingspatroon is o.a. gebaseerd op de vereiste quality of services en de risicobereidheid van de vrager.

De intake van een vraag voor een informatieproduct volgt globaal drie fasen:

1. Inventariseren

Vanuit de behoefte voor een nieuw of aan te passen informatieproduct dan wel de incentive van een te mitigeren risico wordt een inventarisatie gemaakt van:

- De belangrijkste (niet beïnvloedbare) product- en proceskenmerken;
- (Indien nodig) de issues met huidige productieketen en (IST);
- (Indien nodig) een schets/prototype en/of beschrijving van de gewenste situatie (SOLL).

2. Beoordelen

Op basis van de geïnventariseerde eigenschappen wordt samen met de eindverantwoordelijke (business owner) ingezoomd op:

- De doeltreffendheid en doelmatigheid van het informatieproduct (beïnvloedbare product- en proceskenmerken);
- Productrisico's en kans/impactanalyse[54];
- Uitvoering van een criticality assessment (CA), WMK en/of privacy impact assessment (PIA) (indien nodig).

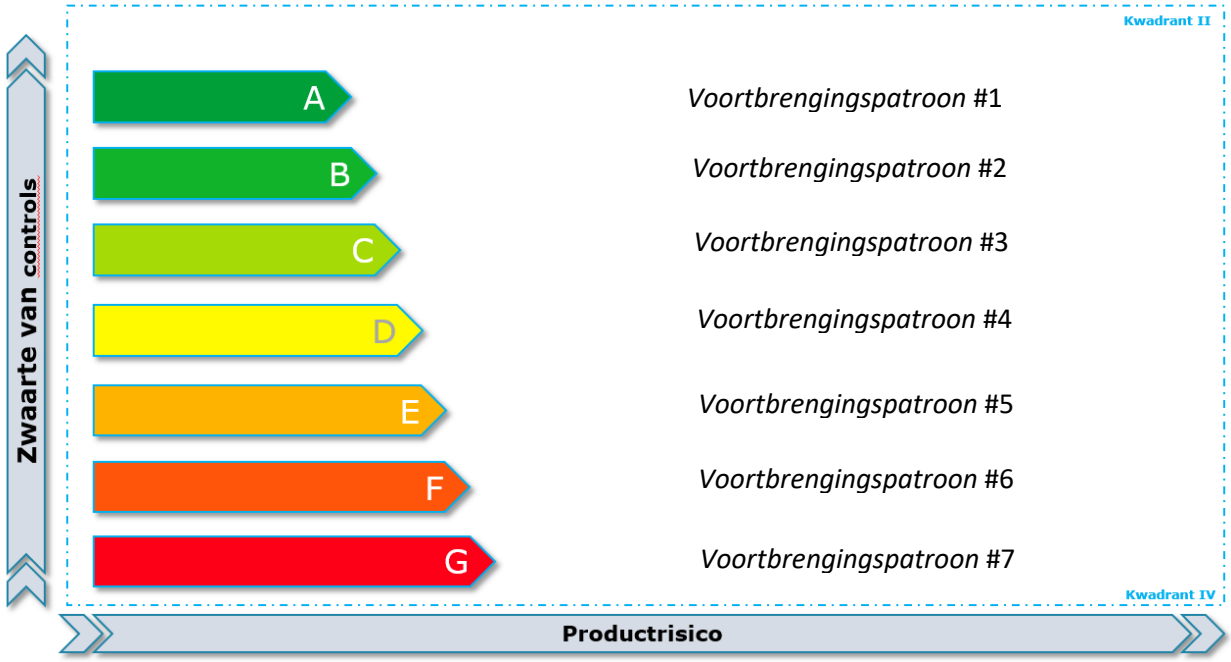
3. Toekennen

Op basis van het totaal aan product- en proceskenmerken en de risico-inventarisatie wordt toegekend:

- Dataclassificatie
- Risicoreactie;
- Duurzaamheidslabel;
- Voortbrengingspatroon (oplossingskader)

Deze drie stappen worden door de domeinen gegevens en analytics altijd samen met de vrager/business owner uitgevoerd.

De solution architect bepaalt op basis van de risicoreactie en de verzamelde productkenmerken het duurzaamheidslabel. Dit label is oplossingskaderstellend wat betekent dat de ketenimplementatie van de productvoortbrenging met de systeemcomponenten en -patronen gerealiseerd dient te worden die in dit oplossingskader zijn opgenomen.



[54] Informatiebeveiligingsrisico's, continuiteitsrisico's, ICT risico's, juridische risico's, HR risico's, process risico's, verslaleggingsrisico's.

Hoeveel voortbrengingspatronen er uiteindelijk zijn/komen moeten de respectievelijke domeinarchitecturen van gegevens en analytics beschrijven. Het duurzaamheidslabel is onderdeel van de metadata van het informatieproduct en is transparant zichtbaar voor de gebruiker van het product.

b) Een polygestructureerde[55] variant[56]

Het gaat hier om een dienst die inzicht geeft in de (op dit moment) 4 miljard bestanden van de Belastingdienst. Uitgaande van 10 datapunten per bestand zijn dat 40 miljard datapunten waar tot op heden een minimale governance op zit. De kans dat er in die 40 miljard datapunten persoonsgebonden data staat (bv. Een kenteken), is aanzienlijk. Om hier grip op te krijgen moet er een dienst gerealiseerd worden die bedrijfsonderdelen in staat stelt:

- om inzicht en overzicht te krijgen wat er allemaal voor polygestructureerde producten op de systemen staan;
- deze producten te kunnen annoteren, archiveren en terug te kunnen vinden
- deze producten op te kunnen ruimen;
- de afname van deze producten in tijd te monitoren en op te kunnen besturen.

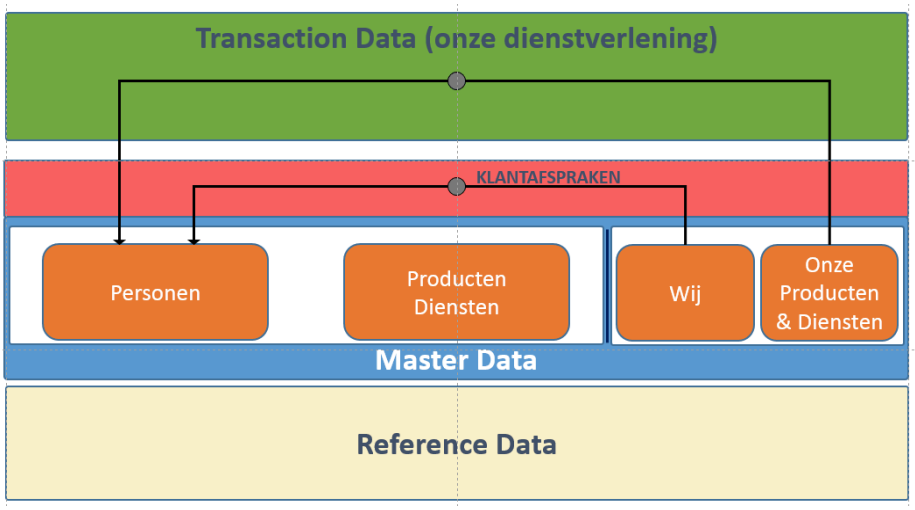
[55] Polygestructureerd verwijst naar documenten of andere vormen van dataopslag als pdf, Excel, email, txt, docx, pptx, etc. Alles behalve data komende uit bronadministraties dan wel gegevensleveringen van derden. 'Ongestructureerd' in dit kader zou een foutieve weerspiegeling zijn van de informatiewaarde die wel degelijk in dit soort documenten/dataopslag aanwezig is

[56] Deze dienst is reeds in prototype gerealiseerd door het cognitieve team bij IV accent

Master- en Referentie data gegevensdienst

Het is van belang om zo duidelijk mogelijk te zijn over de classificatie van typen data. Het type bepaald de gegevensdienst waarin het moet worden opgenomen. En de gegevensdienst wordt gerealiseerd door middel van een technische voorziening (bv. MIH).

Als uitgangpunt voor de classificatie geldt de volgende figuur:



Figuur 3: Classificatie data typering, transactie, master, klantafspraken en referentiedata

Masterdata	In zijn algemeenheid valt de masterdata onder te verdelen in een 'WIJ' en de 'Rest van de wereld'. *Rest van de wereld* Daaronder vallen die 'dingen' die geheel eigenstandig zijn te definiëren en waar slecht één partij in vrije wil over de levenscyclus kan besluiten. Een persoon kan besluiten zijn geslacht te veranderen, te trouwen en de achternaam van de partner over te nemen, etc.. Andere 'dingen' zijn producten en diensten die ook geheel eigenstandig zijn te definiëren. Denk aan een woning, auto, etc.. Bovenste twee categorieën van data zijn 'dingen' waar de Belastingdienst DUS geen regie op heeft, maar die gegevens wel nodig heeft. Dit noemen we *system of reference*. *Wij* Belastingdienst eigen gegevens, 'WIJ' die onszelf definiëren (en relateren aan die anderen). Onze producten en diensten zijn: inkomensheffing, OB heffing, Erf belasting, etc. Hiervan is de Belastingdienst dus de *System of record.*

Klantafspraken	Klantafspraken zijn afspraken tussen 'WIJ', de Belastingdienst, en de 'PERSONEN'. Bijvoorbeeld, 'Wij', de Belastingdienst, hebben afspraken met 'PERSOON X' dat hij een ondernemer is en dus OB (product/dienst) plichtig is per kwartaal. Ook wel; middel-gerelateerde gegevens; klantafspraken, wij 'wijzen toe' aan anderen ten behoeve van transacties).
Transactiedata	Als wij een van onze producten aan anderen leveren of diensten aan anderen verlenen dan doen wij een transactie. Alles wat we daarvan administreren zijn transactie-gerelateerde gegevens.

Masterdata en Klantafspraken worden beiden door de master- en referentie gegevensdienst geleverd!

Voorbeeld:

Als wij een kenmerk aan een andere persoon 'toewijzen' (op basis van onze regels) dan is dat dus niet direct 'master data'. Dat zijn afspraken die twee partijen maken op basis van gedeelde regels met een levenscyclus die meerdere (vele) transacties overspannen. Het zijn wel persoon-gerelateerde gegevens en in dat opzicht dicht tegen master-data aanschurend maar niet door de persoon zelf in vrije wil bepaald, er zit een gedefinieerd 'samen' element in.

De classificatie is dat dat het 'Klantafspraken' zijn, ook wel middel-gerelateerde klantgegevens. Deze data willen we logisch dicht tegen het persoon aanhangen, de ge-eigende gegevensdienst hiervoor is de Master- en referentie data gegevensdienst.

Voor alle concepten die domeinoverstijgend zijn en waar de instantiaties een zogenaamd master record hebben, moet een centrale master- en referentie gegevensdienst worden opgezet, deze dienst is:

- De centrale system of reference voor de bronadministraties t.a.v. concepten waar de Belastingdienst niet de system of record van is (bv. BRP gegevens t.a.v. burgers, KVK gegevens t.a.v. bedrijven maar ook bv. landencodes die van ISO komen);
- De centrale system of record voor de bronadministraties t.a.v. concepten waar harmonisatie en eenmalige administratie cruciaal is (bv. de fiscale context van een subject of subject gebaseerde gegevens zoals een postadres).

Randvoorwaarden:

- De dienst levert services die interne sleutels uitlevert aan de bronadministraties t.b.v. sleutelintegratie over bronnen, processen en domeinen heen. Deze sleutels mogen nooit naar buiten gecommuniceerd worden;

- De dienst levert matching en merging functionaliteit t.a.v. de instantiaties van concepten;
- De dienst levert services CRUD services t.a.v. het onderhoud van system of record concepten aangaande de in de master- en referentiedata opgenomen gegevens.

Concrete invulling t.a.v. persoonsgegevens:

- Gegevens over Persoon of bedrijf ontvangt de Belastingdienst uit basisregistraties, de Belastingdienst heeft van deze data dus de *system of reference;*
- Voor de fiscale verplichting en bijbehorende (vaak fiscaal gerelateerde) gegevens die hangen aan een persoon of bedrijf is de Belastingdienst de system of record, dit noemen we ook wel de middelspecifieke klantgegevens;
- Voor beide type gegevens is de master- en referentie data gegevensdienst[57] de ge-eigende voorziening[58].

Met de master- en referentie data gegevensdienst wordt voorkomen:

- dat er steeds weer domein specifieke administraties gemaakt en onderhouden moeten worden voor subject gerelateerde fiscale gegevens;
- dat er een woud aan kopieën en koppelingen gaan ontstaan;

Daarmee wordt deze data ook domeinoverstijgend ter beschikking gesteld, bijvoorbeeld t.b.v. Toezicht en Interactie.

Let wel; er wordt hier een duidelijk onderscheid gemaakt tussen de verantwoordelijkheid van de data, die blijft bij het betreffende domein en het domein die de dienst levert (in dit geval domein gegevens).

[58] Voorbeelden zijn de WGA (LH) en de OB klantregistratie

6. Belastingdienstbrede belangen

Modelautoriteit

Het succes van de gegevensdiensten berust op de organisatorische discipline die de organisatiebrede belangen behartigen. Het is volstrekt begrijpelijk dat domeinen hun eigen belangen najagen, maar m.n. m.b.t. (meta)data/masterdata en services[59] moeten organisatiebrede belangen expliciet belegd worden.

Er moet daarom een soort van *autoriteit (of autoriteiten)* worden ingericht. Let wel, de domeinen moeten zelf in staat worden gebracht om de verschillende ontwerp artefacten te maken. De autoriteit toetst expliciet op de organisatiebrede belangen, denk aan een vier-ogen-principe, maar ook een consultancy functie die de domeinen helpt de ontwerp artefacten fehig te maken voor de gehele organisatie.

Deze autoriteit heeft de volgende verantwoordelijkheden:

T.a.v. (meta)data/masterdata:

- Voldoen informatie- en datamodellen aan de *domeinoverstijgende belangen*, denk aan:
  - Naamgevingen, definities, identificaties;
  - Hergebruik van concepten en instantiaties;
  - Is de datadefinitie architectuur / model catalogus kwalitatief voldoende.
- Voldoen informatie- en datamodellen aan de eisen t.a.v. *deploybaarheid*, denk aan:
  - Methodologische correctheid van de informatie- en datamodellen;
  - Een traceerbare, af te leiden en uiteindelijk te automatiseren modeltransformatie van hogere niveaus van representatie naar lagere niveaus.
- Zijn de informatie- en datamodellen voldoende voorzien van de noodzakelijke metadata, denk aan:
  - Dataclassificatie;
  - Gegevensverantwoordelijke;
  - Doelbinding;
  - Belangen (Te Beschermen Belang, Privacy Belang, Cruciaal Belang, Archief Belang, Strafrechtelijk Belang);
  - Etc.
- Voldoet de technische implementatie van data aan de eisen die worden gesteld aan de voortbrenging richting set gebaseerde gegevensdiensten.

[59] Denk aan Application Programming Interfaces, m.n. de subject gebaseerde gegevensdienst zal bestaan uit veelvoud van API endpoints.

T.a.v. Services

De subject gebaseerde gegevensdienst bestaat uiteindelijk uit vele services die aangeroepen kunnen worden. Bijvoorbeeld een composite/business service die op basis van een BSN#, andere meer elementaire services aanroept, die uit verschillende administraties een openstaande vordering haalt. De compositeservice ontvangt de elementaire responses en voegt deze samen in een structuur die gevraagd is.

Voorkomen moet worden dat er een wirwar van services gaan ontstaan die veel gaan overlappen en waar het belang van hergebruik niet wordt onderkend. Het onderhouden van een service catalogus met daarin goed geadministeerde metadata die publiceerbaar en doorzoekbaar is, is essentieel.

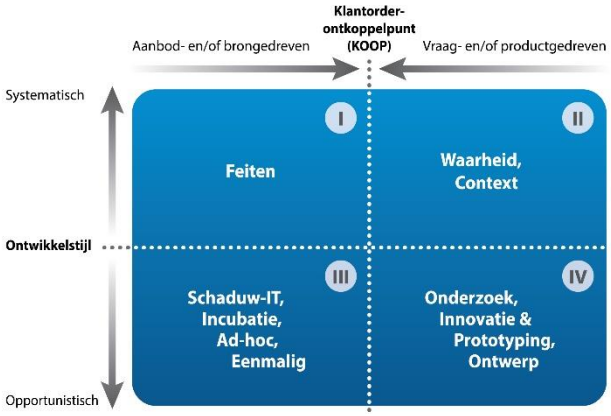
De service autoriteit moet deze belangen gaan afdekken.

Consultancydiensten t.a.v. gegevens

Met name aan modelleringskant, de zogenaamde datadefinitie architectuur, moet veel gebeuren t.a.v. opleidingen en consultancy. Bestaande modelleringspraktijken moeten unlearned worden en nieuwe moeten worden geleerd. Denk aan feit gebaseerd modelleren, logisch modelleren en anchor stijl modelleren. m

7. Data Kwadranten Model, scenario's en afbakening

Voor de afbakening van het domein gegevens wordt het zogenaamde data kwadrantenmodel[60] gebruikt. Voor een gedetailleerde beschrijving van het model wordt verwezen naar bijlage A van dit document.



Figuur 4: Data Kwadranten Model (Damhof)

Het model is een hulpmiddel voor bestuurders, managers, specialisten en domein experts om over het gegevensdomein te discussiëren, beleid en strategie te formuleren, afbakeningsdiscussies zuiver te voeren en vooral een gevoel te ontwikkelen aangaande de samenhang van gegevensdiensten en de samenhang met analytics.

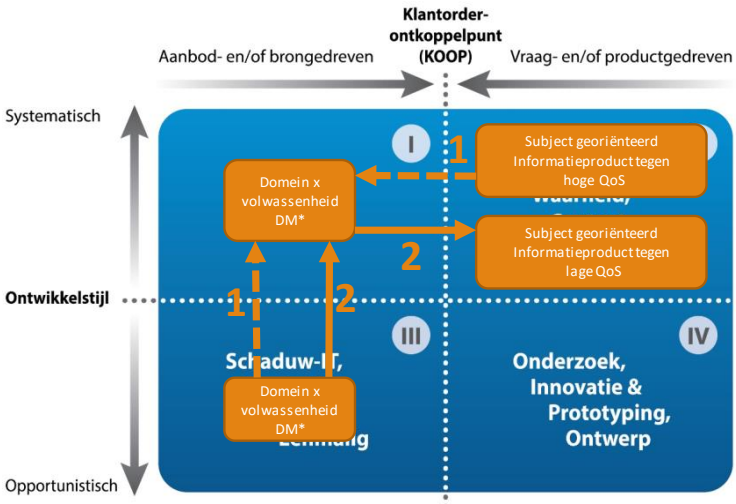
Belangrijk uitgangspunt is dat alle data die de Belastingdienst tot haar beschikking heeft op het kwadrantenmodel te plotten is. Elke system of record, gegevenslevering van en aan derden, Excel bestand, datafundament, pdf, DSS of Enterprise Data Warehouse heeft een positie in het kwadrant.

[60] Zie ook Tijdschrift voor Toezicht, september 2018; Opzet van een datavoorzieningsfunctie ter ondersteuning van datagedreven toezicht, Damhof/Aerle/Ouddeken, een extract hieruit is in bijlage A opgenomen.

Rootcause scenario, Data Kwadranten Model

In hoofdstuk 0 is de volgende root-cause beschreven:

> *De proliferatie van gegevens is voor het grootste deel een manifestatie van 'dataschuld[61]' in de bronadministraties.*



Figuur 5: root-cause t.a.v. dataschuld in bronadministraties

Deze root-cause speelt zich als volgt af, gevisualiseerd met behulp van het Data Kwadranten Model:

Uitgangspunt:	#1. Een bronadministratie X heeft een lage datavolwassenheid[62] (positie in kwadrant III)[63]. #2. Het gewenste informatieproduct in kwadrant II moet aan hoge quality of services voldoen. #3. Kwadrant II informatieproducten mogen alleen uit kwadrant I komen.
Scenario #1: *Pull*	Breng de bronnen in middelendomein X naar een hoger datavolwassenheidsniveau (naar kwadrant I), het informatieproduct kan eenvoudig realtime afgeleid worden (tegen hoge quality of services). Bv. m.b.v. API's (pull van gegevens, eventueel middels technische replicatie/mirror van bron).
Scenario #2: *Push*	Ontsluit de data uit de bron en werk deze op met de producten en diensten van het Domein Gegevens, deze opwerking betreft dus het kopiëren, integreren, uniformeren en transformeren van data (push van gegevens). Het domein gegevens MOET dus de gegevens van het betreffende domein kwadrant I waardig maken.

[61] Naar analogie van de definitie van technische schuld: De schuld bestaat uit de kosten voor de herstelwerkzaamheden (Engels: refactoring) die uitgevoerd moeten worden om een consistente en onderhoudbare oplossing te realiseren. Zo lang dit werk niet gedaan is blijft de schuld bestaan en wordt er telkens rente betaald, in de vorm van extra inspanning voor wijzigingen.

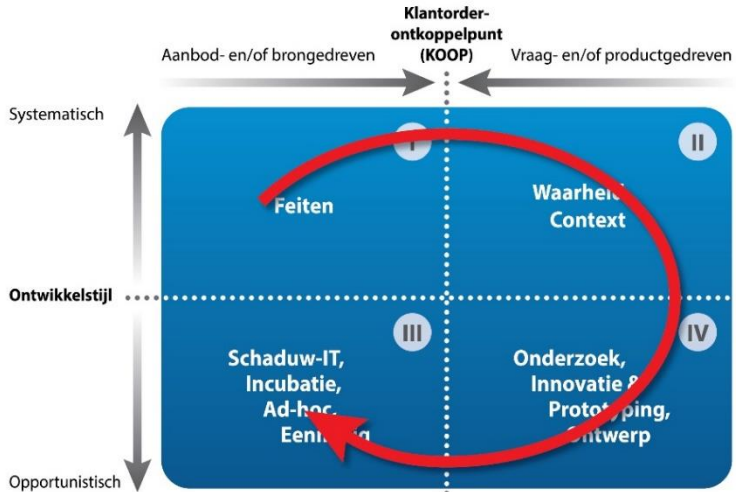
[62] Datavolwassenheid kan worden vastgesteld met de Data Management Maturity Model van CMMI, uitgebreid met specifieke BD kaders.

[63] Bv. de data conformeert zich niet aan de feit- en/of logische datamodellen, identificatie van objecten/subjecten is niet geharmoniseerd, etc..

Het betreffende informatieproduct heeft een lagere QoS (!); latency en opwerkingscomplexiteit veroorzaken dat. **DIT IS NIET OP TE LOSSEN.**

Keuze in organisatorische inrichting en governance

Een referentiearchitectuur geeft geen, of minimaal, advies t.a.v. de wijze waarop de Belastingdienst zich moet organiseren t.a.v. gegevens. Echter, de architectuur geeft wel handvaten t.a.v. de variabelen die de organisatievorm bepalen. Zo kan de mate van centralisatie versus decentralisatie geplot worden op het Data Kwadranten Model:



Figuur 6: De mate van centralisatie (linksboven) vs. de mate van decentralisatie (linksonder)

Centralisatie

Een groot deel van de systematische voortbrenging (kwadrant I en deel van II) heeft een tendens richting centralisatie, met name omdat:

- Er belastingdienstbrede belangen worden afgedekt. Denk aan domeinoverstijgende data-integratie t.a.v. vele verschillende gebruikerstyperingen (analytics, management informatie, apps, CRUD, etc.) t.b.v. vele verschillende doelgroepen (burger, bedrijf, medewerker, ketenpartners, etc.), rekening houdende met data-aspecten die belastingdienstbreed noodzakelijk worden geacht (privacy by design, tijdreizen, traceerbaarheid, etc.);
- De quality of services m.n. zich manifesteren in duurzaamheid. Denk daarbij aan robuustheid, onderhoudbaarheid, veranderbaarheid, etc.
- Er sprake is van ver doorgevoerde functiescheiding t.a.v. ontwerp, realisatie en beheer van zowel datastromen en opwerkingen, applicaties en infrastructuur;
- Er schaarste is van kennis en competenties waardoor centralisatie efficiënt is;
- Er investeringen nodig zijn in middelen die waar synergie maximaal moet worden uitgenut(lees; organisatiebreed).

Decentralisatie

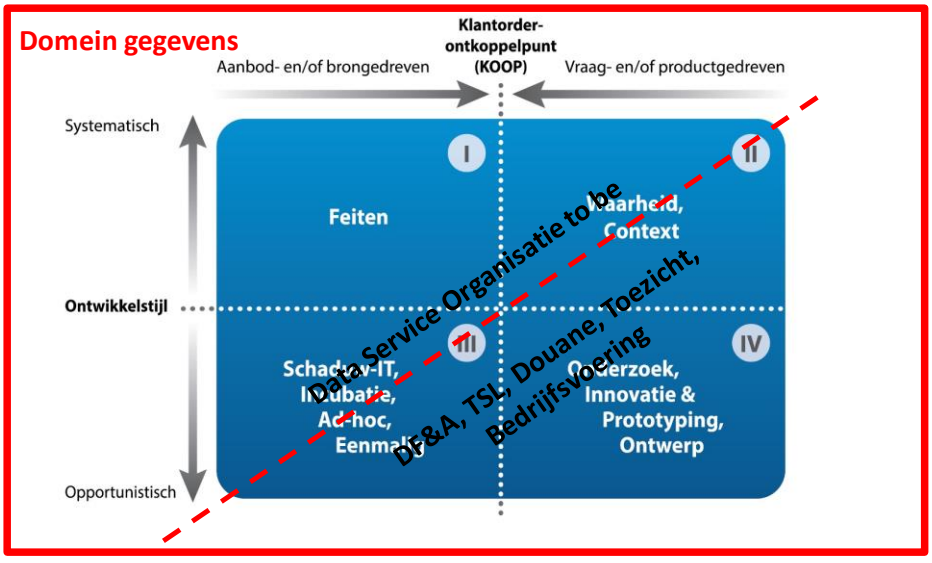
Een groot deel van de opportunistische voortbrenging van gegevens (kwadrant III en IV) heeft een neiging tot decentralisatie naar verschillende domeinen, bedrijfsonderdelen, teams en zelfs individuele medewerkers, met name omdat:

- Kennis van het specifieke domein en de data van doorslaggevend belang is om het betreffende informatieproduct te kunnen maken;
- De wijze van voortbrenging zich kenmerkt door iteraties, experimenteren, exploreren en ontdekken;
- Er minimale functiescheiding is t.a.v. datapreparatie, maken en testen analytisch model- en of rapportage ontwikkeling en interpreteren van de resultaten;
- Er ruimte moet worden gehouden om data op te werken naar eigen inzichten.

De constatering is dat binnen de Belastingdienst nagenoeg alle domeinen zich bezig houden met kwadrant III en kwadrant IV voortbrenging.

Domein gegevens versus bedrijfsonderdelen

Alle data en datastromen vallen onder het domein gegevens en de daarvoor opgestelde bedrijfsonderdelen, echter afhankelijk van de risicobereidheid en het duurzaamheidslabel van het gevraagde informatieproduct[64] mogen andere bedrijfsonderdelen ook data voortbrengen.



Figuur 7: Domein gegevens versus bedrijfsonderdelen

[64] Een managementrapportage, analyse t.b.v. tweede-kamer vraag, eenmalig beleid beïnvloedende analyse op profiel zzp die frauderen, tot een geavanceerd analytics model, risico detectie service of dataset t.b.v. apps of leveringen aan instituties

Ofwel:

a) Informatieproducten met een lage risicobereidheid en hoog duurzaamheidslabel worden qua data geheel door het domein gegevens voortgebracht (Kwadrant I/II);
b) Informatieproducten met een hoge risicobereidheid en laag duurzaamheidslabel kunnen deels wordt voorgebracht door andere dienstonderdelen, o.a. DF&A (kwadrant III/IV);
c) Informatieproducten die gepromoveerd moeten worden naar een lagere risicobereidheid (a.k.a. hoger duurzaamheidslabel) doorlopen een promotietraject waarin het domein analytics en het domein gegevens samen optrekken.

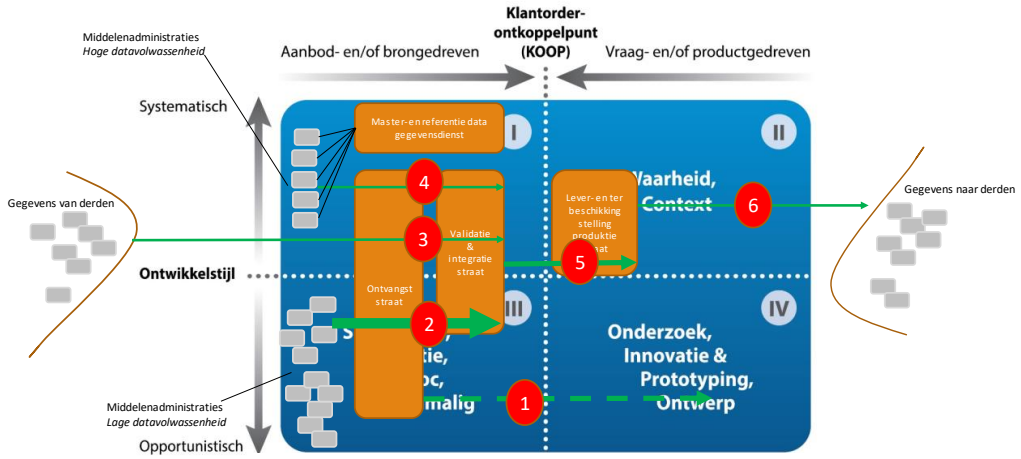
'Data Service Organisatie' als bedrijfsonderdeel bestaat niet en is een verzamelnaam voor CAP en verschillende IV onderdelen die zich bezighouden met gegevens. De TVB nota[65] is een soort eerste fase in een ontwikkeling richting een dergelijke Data Service Organisatie.

Belangrijk is aan te geven dat de domeinindeling niks zegt over de wijze waarop de uitvoering uiteindelijk wordt belegd. Bovenstaande is dus geschreven vanuit een IST t.a.v. de wijze waarop de uitvoering is georganiseerd.

[65] Memo Taken, verantwoordelijkheden, bevoegdheden bronontsluiting en interne beschikbaarstelling gegeven, oktober 2018

Domein gegevens en de set-gebaseerde gegevensdiensten

De set-gebaseerde gegevensdiensten (a.k.a. de Belastingdienst Data Fabriek) zoals beschreven in hoofdstuk 5 zijn in onderstaand figuur afgebeeld. De figuur geeft een visuele representatie hoe de set-gebaseerde gegevensdienst zou werken in het hier en nu.



Figuur 8: set-gebaseerde gegevensdiensten in het hier en nu

Alle data van de bronadministraties en gegevensleveringen van derden worden ter beschikking gesteld van de ontvangststraat. Globaal zijn daarin de volgende datastromen een optie:

1	De data uit alle bronadministraties is beschikbaar (mits toegestaan) om in kwadrant IV mee te experimenteren. De data is as-is zoals de aanleverende bron dat ter beschikking stelt. Ook eigen data[66] kan ter beschikking worden gesteld.
2	Bronadministraties met een lage volwassenheid moet worden opgewerkt en compliant worden gemaakt aan een expliciet te modelleren datamodel en wordt daarop ook gevalideerd. *Dit is een relatief dikke pijl omdat er van wordt uitgegaan dat er veel 'dataschuld' opgelost moet worden.*
3	Uitgangspunt is dat bij het merendeel van gegevensleveringen van derden de vrager de eisen kan stellen t.a.v. de wijze en kwaliteit van aanlevering. Allerlei concerns m.b.t. validatie en integratie zijn daarom wel van belang maar relatief eenvoudig (dunne pijl)
4	Bronadministraties met een hoge datavolwassenheid vereisen minimaal tot geen validaties (ze zijn ten slotte compliant met een gepubliceerd feiten- en of logisch model) en ook de sleutelintegratie is relatief eenvoudig (dunnen pijl)
5	Gevalideerde en geïntegreerde data kan op verschillende wijzen (multirealiteiten) worden uitgeleverd (bv. bij fysieke leveringen) of ter beschikking worden gesteld.
6	Het fysiek verzenden of toegankelijk maken van gegevens t.b.v. derden

[66] Dit is eigen data van een analist of datascientist die deze data wil gebruiken i.c.m. andere (bv. brinadministratie) data in kwadrant IV.

De set-gebaseerde diensten leveren dus ondersteuning voor de kwadrant I en II voortbrenging van gegevens waarbij het inwinnen van gegevens en verstrekken van gegevens van en naar derden onderdeel is van de scope.

Het uitgangspunt van de dienst is dat data zoveel mogelijk *ter beschikking wordt gesteld* i.p.v. fysiek uitgeleverd.

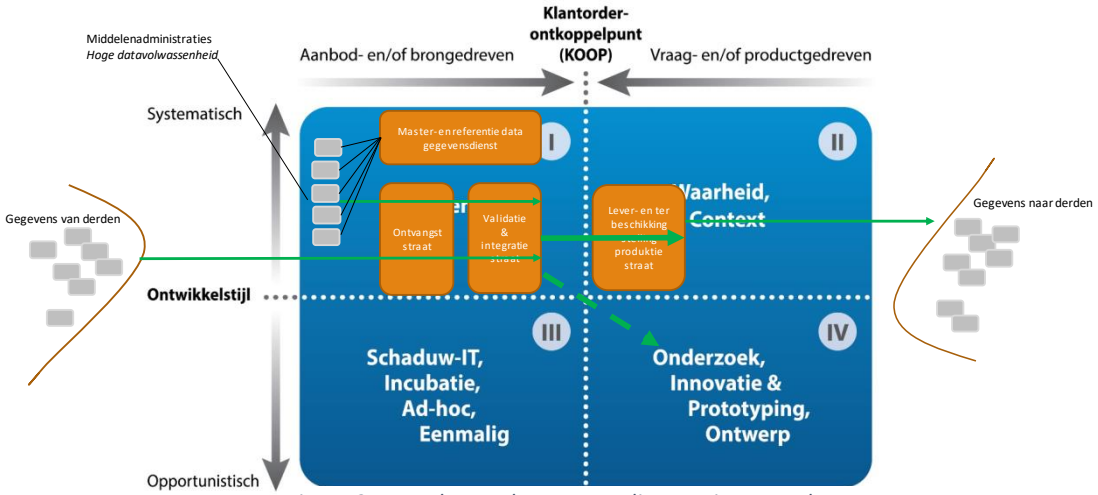
De set-gebaseerde diensten geven dus een hoogwaardig alternatief voor diensten die voordien door het Enterprise Data Warehouse werden uitgevoerd.

De set-gebaseerde diensten zijn een cruciale dienst om opwerking en integratie van data mogelijk te blijven maken in een transitie waar de bronadministraties geleidelijk naar een hogere datavolwassenheid groeien.

De set-gebaseerde diensten worden significant minder complex (en dus beter beheersbaar, veranderbaar en onderhoudbaar) als de bronadministraties een hogere datavolwassenheid bereiken.

De ontvangststraat data mag door domeinen (mits toegestaan) buiten de fabriek verder worden opgewerkt (bv. t.b.v. analytics).

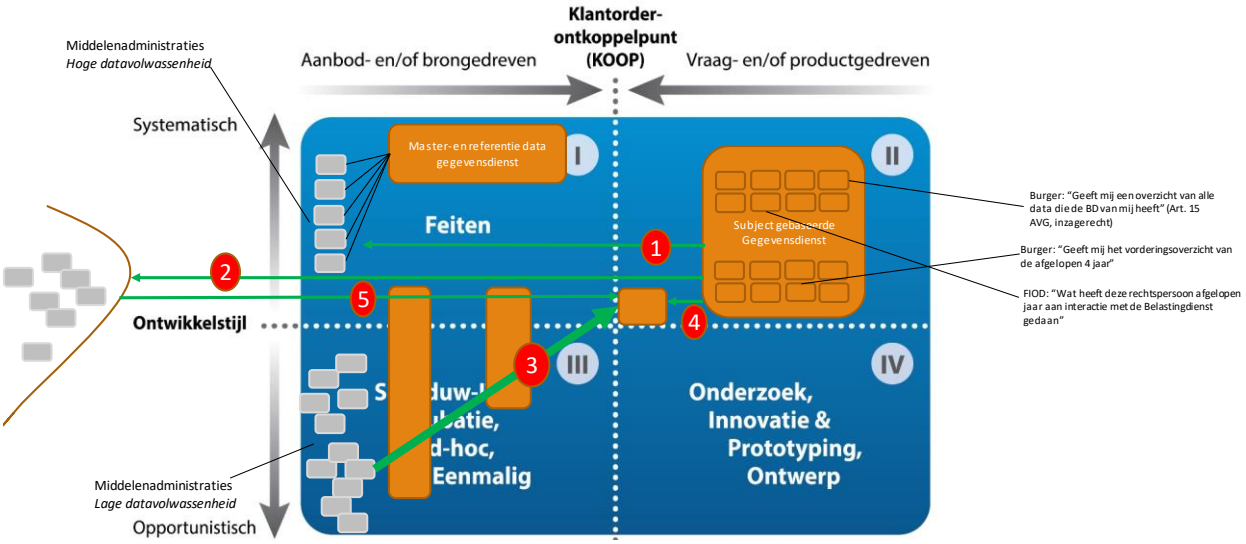
In een ideaal scenario met alleen maar authentieke bronnen (binnen en buiten de Belastingdienst) die gekenmerkt worden door een hoge datavolwassenheid, ziet e.e.a. er als volgt uit:



Figuur 9: set-gebaseerde gegevensdiensten in een end-state

Domein gegevens en de subject-gebaseerde gegevensdiensten

De subject-gebaseerde gegevensdiensten zoals beschreven in hoofdstuk 5 zijn in onderstaand figuur afgebeeld.



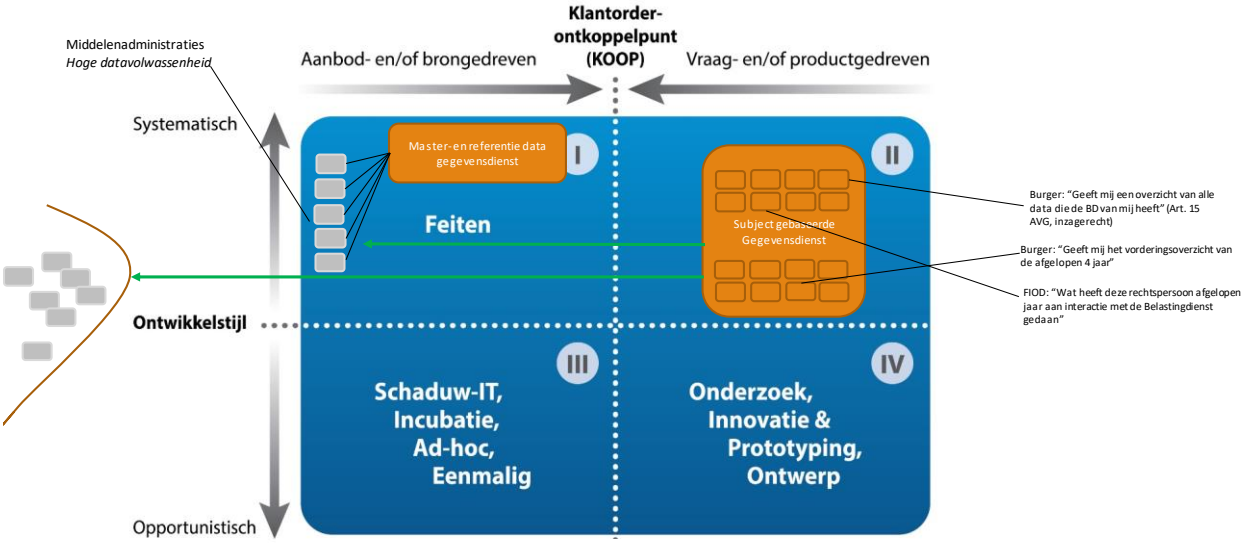
Figuur 10: subject-gebaseerde gegevensdiensten in het hier en nu

De subject-gebaseerde diensten stelen medewerkers, burger en bedrijf in staat om op basis van een sleutel (bv. BSN#) een actueel en relevante informatiepositie op te vragen, te kunnen valideren en te wijzigen.

1	(pull) Happy flow scenario, de gestelde subject-gerichte vraag kan worden beantwoord door het uitvragen van authentieke bronnen met een hoge volwassenheid, geen opwerking, validatie of integratie van gegevens nodig.
2	(pull) Ook de ketenpartners hebben subject gerichte diensten. Gevraagde data waar de authentieke bron bij de ketenpartners staat kan worden verkregen op het moment dat het nodig is.
3	(push) Data binnen de Belastingdienst die wordt gekenmerkt door lage datavolwassenheid, maar toch nodig is voor subject-gebaseerde gegevensdienst moet worden ontsloten, gevalideerd, geïntegreerd alvorens deze geconsumeerd kan worden.
4	(pull) Als data uit authentieke bron is ontsloten, gevalideerd en geïntegreerd kan deze door de subject-gebaseerde dienst geconsumeerd worden. De quality of service kenmerkt zich altijd door latency (niet actueel) en datakwaliteit en proces risico's.
5	(psuh) Voor die ketenpartners die nog geen subject-gebaseerde gegevensdiensten hebben moet de data eerst worden ingewonnen, gevalideerd, geïntegreerd en ter beschikking worden gesteld.

De subject-gebaseerde gegevensdiensten hebben in een transitie altijd een afhankelijkheid van de set-gebaseerde gegevensdiensten.

In een ideaal scenario met alleen maar authentieke bronnen (binnen en buiten de Belastingdienst) die gekenmerkt worden door een hoge datavolwassenheid, ziet e.e.a. er als volgt uit:

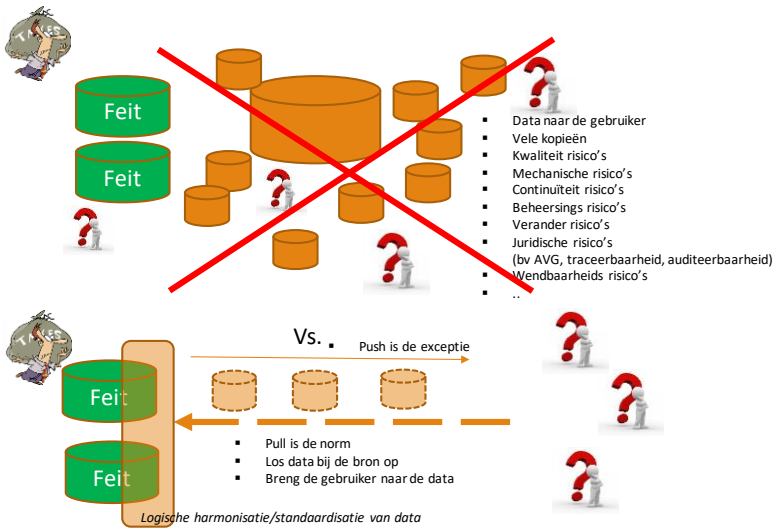


Figuur 11: subject-gebaseerde gegevensdiensten in een end-state

8. Architectuurkaders gegevens

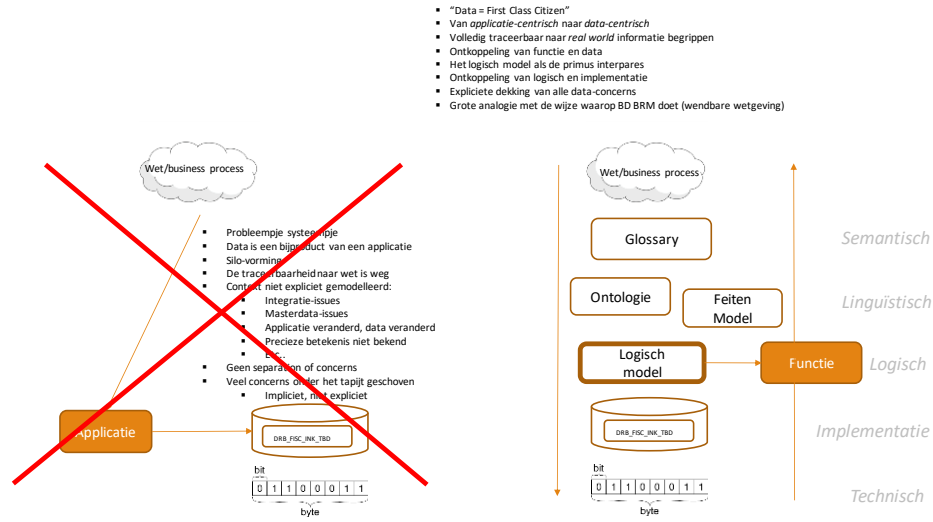
Elke data-architectuur kent twee hoofd-perspectieven;

1. De horizontale data-architectuur, ook wel de data-logistieke architectuur gaat over de wijze waarop data zich beweegt van A naar B (flow);



Figuur 12: De horizontale data-architectuur

2. De verticale data-architectuur, ook wel de datadefinitie architectuur verwijst hoe een concept in een wettekst of uit de reële wereld zich uiteindelijk vertaald in een fysieke implementatie (veelal een database). Het gaat dan ook om de data-architectuur van de data-in-rust (state).



Figuur 13: De verticale data-architectuur

Bij beide perspectieven op data-architectuur staan de 'separations[67] of concerns' centraal.

[67] Edsger Dijkstra, 1974, "On the role of scientific thought"

Leeswijzer

De Data Definitie Architectuur (DDA) is relevant voor alle middelen-administraties en andere vormen van feiten-vastlegging. De Data Logistieke Architectuur (DLA) is relevant voor alle vormen van opwerking en deling van gegevens met andere domeinen. De Data Governance (DGO) kaders zijn met name gericht op afbakenings- en organisatorische kaders. Verder is er een kolom D opgenomen die staat voor de relevantie van het betreffende domein; 'A' staat voor Alle Domeinen, 'G' staat voor Gegevens Domein.

<<Rationale en implicaties per kader zijn nog niet beschreven, zouden wel afgeleid/verklaard moeten kunnen worden uit de hoofdstukken 1 tm 6>>

Sectie	D	#	Beschrijving
(DDA) Data Definitie Architectuur	A	1[68]	Alle state van data is gemodelleerd. Zo veel mogelijk concerns/belangen m.b.t. data worden expliciet gemodelleerd , echter verschillende concerns worden door verschillende modelleringsmethodieken afgedekt (niveaus van representatie; semantisch, linguïstisch, logisch, implementatie). De wijze van modelleren is per niveau van representatie gestandaardiseerd. Modellen representeren altijd de daadwerkelijke situatie in productie.
	A	2	Logica (Regelmodellen[69]) en datamodellen (data is altijd een vorm van logica) moeten gevalideerd kunnen worden door domein experts op basis van natuurlijke taal en voorbeelden.
	A	3	Semantische concerns (verantwoordelijke[74], definities, relatie naar wetteksten, jurisprudentie, etc..) t.a.v. concepten/bedrijfsobjecten worden bijgehouden in glossary/woordenboek achtige gereedschappen door de respectievelijke verantwoordelijken van die bedrijfsobjecten/concepten. Domein gegevens beheert de glossary en bepaalt de tooling, ieder domein heeft een eigen verantwoordelijkheid voor haar de content van de glossary. *De gegevensgebiedenarchitectuur[70]geeft een overzicht van bedrijfsobjecten/concepten en de verantwoordelijke partij t.a.v. definities en relaties. Dit document is kaderstellend t.a.v. modelleringen van bronadministraties.*
	A	4	Linguïstische concerns worden gemodelleerd in ontologieën (t.a.v. classificatie concerns) en feitgebaseerde[71] informatiemodellen (t.a.v. communicatie, verbalisatie en validatie concerns).

[68] Er worden niveaus van representatie onderscheiden zowel in de Data Definitie Architectuur als in de Data Logistieke Architectuur, uitgangspunt is daarbij de concernmatrix in bijlage H.

[69] Deze praktijk is al gaande bij BRM (Business Rule Management)

[70] Gegevensgebieden en administraties (model: gegevensgebiedenarchitectuur) geeft een hoofdordening voor bedrijfsobjecten/concepten en wijst verantwoordelijken aan t.a.v. definitie (semantische concerns) en relaties (linguïstische concerns)

[71] Deze vorm van modelleren voldoet aan de eisen t.a.v. theorie/methodologie, aanpak en notatie. Het kent zijn oorsprong in NIAM van Sjir Nijssen en is onder meer doorontwikkeld als CogNIAM en FCO-IM op de Radbout Universiteit Nijmegen. Ook Object Role Modeling (Terry Halpin) komt uit deze stroming. De vorm van modelleren wordt gebruikt in DEMO en is ook gebruikt in de POC gegevens bij kwadrant IV waarmee goede ervaringen zijn opgedaan. Belangrijk is dat feitgebaseerd modelleren de zogenaamde UoD (Universe of Discourse) als uitgangspunt neemt, denk aan de wet, een domein of een gegevensgebied. In elk geval niet de applicatie.

			Gebruik van concepten kan alleen als het betreffend concept verwoord is in de glossary/woordenboek [DDA.3]
	A	5	Logische concerns (integriteit, manipulatie, afleiding, consistentie, etc..) worden gemodelleerd in zogenaamde logische modellen welke de Relational Data Model principes volgen (zie bijlage B).
	A	6	De volgende (oa AVG) kenmerken worden op (minimaal) entiteit/attribuut (niet rij) niveau vastgelegd (zie bijlage E); bewaartermijn, labeling van belangen[72], BIV classificatie[73], gegevensverantwoordelijke[74], doelbinding , grondslag en 'te pseudonimiseren' attributen.
	A	7	T.b.v. operationalisatie van datamodellen is het logische model voorwaardelijk en het feitenmodel een pre.
	A	8	Alle operationalisaties van datamodellen (deployment) genereert zich (al dan niet geautomatiseerd) uit een hoger niveau van representatie, minimaal uit een logisch model. *NB. De operationalisatie-strategie van een feiten- en of logisch model wordt afgestemd aan de hand van de implementatie concerns (performance, schaal, etc..). De technische manifestatie van een logisch model kan vele vormen aannemen, belangrijk is dat het feit en/of logisch model ge-enforced wordt*.
	A	9	Het is niet toegestaan om in productie database artefacten te creëren waar geen logisch en/of feitgebaseerd model aan ten grondslag ligt.
	G	10	Alle niveaus van representaties (implementatie → logisch → linguïstisch → semantisch) zijn maakbaar, volgbaar, geversioneerd, gepubliceerd, geannoteerd (bv. eigenaar, etc.) en onderhouden. Hiertoe moeten gereedschappen ter beschikking worden gesteld
	G	11	Alle verticale modeltransformaties (de niveau van representaties afdalende) zijn zoveel mogelijk geautomatiseerd in patronen.
	A	12	Data die als masterdata en /of klantafspraken zijn gedefinieerd worden gerealiseerd door de master- en referentiedata gegevensdienst. Alle instantiaties van domeinoverstijgende concepten (bv. subject) worden in de bronadministaties geïdentificeerd met een alleen binnen de Belastingdienst te gebruiken intenr nummer waar de Belastingdienst volledige regie over heeft. Hiervoor moet de Master- en referentie gegevensdiensten geschikt worden gemaakt[75]. *Bronadministraties hebben een verplichte winkelnering bij deze dienst.* Speciale attentie t.a.v. het subject en de fiscale kenmerken van dat subject:

[72] Te Beschermen Belang, Privacy Belang, Cruciaal Belang, Archief Belang, Strafrechtelijk Belang.
[73] Zie RA Informatiebeveiliging.
[74] Voorkomen moet worden data gegevenseigenaar een term gaat worden. Eigenaarschap van gegevens is een moeizame discussie en juridisch lastig. Er valt tenslotte wat voor te zeggen dat uiteindelijk de burger of het bedrijf eigenaar is van haar gegevens. Het gaat hier om de verantwoordelijke binnen de Belastingdienst, wie is aanspreekbaar op de gegevens, bv. de betekenis/definitie en de verwerking van gegevens.
[75] De technische voorziening hieronder is de IBM MDM Stack, ook wel MIH genoemd.

			▪ Gegevens over persoon of bedrijf ontvangt de Belastingdienst uit basisregistraties. De Belastingdienst heeft van deze data dus de system of reference; ▪ Voor de fiscale verplichting en bijbehorende (vaak fiscaal gerelateerde) gegevens die hangen aan een persoon of bedrijf is de Belastingdienst de system of record, dit noemen we ook wel de middelspecifieke klantgegevens; ▪ Voor beide type gegevens is de master- en referentie data gegevensdienst[76] de ge-eigende voorziening[77]; ▪ Daarmee wordt voorkomen dat er steeds weer domein specifieke administraties gemaakt en onderhouden moeten worden voor subject gerelateerde fiscale gegevens; ▪ Daarmee wordt voorkomen dat er een woud aan kopieën en koppelingen gaan ontstaan; ▪ Daarmee wordt deze data ook domeinoverstijgend ter beschikking gesteld, bijvoorbeeld t.b.v. Toezicht en Interactie. Let wel; er wordt hier een duidelijk onderscheid gemaakt tussen de verantwoordelijkheid van de data, die blijft bij het betreffende domein en het domein die de dienst levert (in dit geval domein gegevens). *Voorbeeld; De werkgeversadministratie t.b.v. LH valt onder het domein LH, maar moet worden gerealiseerd met de Masterdata gegevensdienst. Samen met domein gegevens wordt deze ingericht. De gegevens zelf blijven onder verantwoordelijkheid van LH vallen.*
(DLA) Data Logistieke Architectuur	A	1[68]	Er worden maximaal zes datalagen onderscheiden t.a.v. horizontale data architectuur, iedere laag heeft zijn eigen concerns (die m.n. logistiek van aard zijn): 1. Bronlaag (system of record); 2. Leveringsabstractielaag/gegevenslaag (techn. DK en pseudonimisering); 3. Logische validatielaag; 4. Centrale (sleutelintegratie) feitlaag (single version of facts); 5. Generieke data access laag (toegang, multirealiteiten); 6. Tool laag (bv. SAS VA heeft zijn eigen in-memory opslag technologie). Eigenaarschap: ▪ Laag 1: het betreffende domein waar het systeem/data bij behoort; ▪ Laag 2 t/m. 6: domein gegevens. NB Laag 1 is de relevante laag voor de bronadministraties en andere vastleggingen Laag 1 bestaat uit twee subtyperingen; ▪ een laag waar de mutaties en transacties daadwerkelijk fysiek gerealiseerd worden ▪ een laag[78] waar de replica/CDC staat t.b.v. de set- en subject-gebaseerde gegevensdiensten.

[76] De Master- en referentie data gegevensdienst is bij de Belastingdienst geoperationaliseerd met de MDM stack van IBM, ook wel MIH genoemd.
[77] Voorbeelden zijn de WGA (LH) en de OB klantregistratie
[78] De b) replicatie/cdc bronlaag wordt gerealiseerd met de datahub technologie van IBM, ook wel MIH genoemd.

			Lagen 2 t/m. 6 hebben betrekking op de set-gebaseerde gegevensdienst Laag 1 heeft relevantie hebben voor de subject-gebaseerde gegevensdienst (technische voorziening=MIH).
	A	2	In zijn algemeenheid is het kopiëren van gegevens niet toegestaan tenzij daar technische concerns voor gelden, uitputtend: 1. Een replica om de belasting op het bronsysteem te ontzien; 2. Het neerzetten van dezelfde data in een structuur die geoptimaliseerd is voor een ander type werklast (bv. analytisch); 3. Het neerzetten van dezelfde data, maar waar pseudonimisatie op is toegepast; 4. Het archiveren van gegevens in een daarvoor geoptimaliseerde service; 5. Enige vorm van intermediaire cache nodig is om een gereedschap te laten functioneren; 6. Het vasthouden van backups t.b.v. continuïteit en herstelbaarheid; 7. Het kopiëren van data t.b.v. een opportunistische opwerking, mits data compartimentering door de ontvanger is ingeregeld
	A	3	Alle horizontale datalagen worden geïmplementeerd vanuit minimaal een logisch datamodel en/of hogere niveaus van representatie
	G	4	Alle horizontale modeltransformaties (de niveau van representaties afdalende) zijn zoveel mogelijk geautomatiseerd in patronen (model- en metadata gedreven). Custom-code wordt zoveel mogelijk vermeden.
	G	5	De gegevenslaag dekt het concern van pseudonimisatie en dus eventuele technische datakwaliteit aspecten t.b.v. pseudonimisatie. Het is bovendien de laag waar opvragingen uit bronnen plaatsvinden[79] en opportunistische opwerking van gegevens t.b.v. bijvoorbeeld analytics starten.
	G	6	Het implementatiemodel van de Centrale FeitLaag is gemodelleerd volgens een anchor style (zie bijlage C).
	A	7	Alle vormen van dataleveringen uit bronadministraties, gegevensleveringen van derden dan wel ter beschikking stellingen zijn gebaseerd op gegevensleveringsovereenkomsten (zie bijlage D), minimaal een logisch model, (o.a. AVG) metadata kenmerken en tenslotte een mapping naar de implementatiemodellen of leveringsbestanden.
	G	8	De metadata van de horizontale data architectuur bevat minimaal: de implementatiemodellen (geversioneerd in tijd) van de verschillende lagen, de datalogistieke resultaten van elke elementaire stap, de validatieresultaten, de afspraken en verplichtingen administratie aangaande aan- en uitleveringen alsmede ter beschikkingstellingen en (AVG) verwerkingskenmerken[80].
	A	9	System-to-system koppelingen binnen een domein moeten op de source-laag worden opgelost (bv. 2 bronnen binnen hetzelfde ketenproces).
	A	10	Domeinoverstijgende system-to-system koppelingen zijn een vorm van een gegevenslevering; moet worden vastgelegd en op de quality of service moet gerapporteerd worden. Deze koppelingen spelen zich af middels de leveringsabstractielaag[81].

[79] Voor die opvragingen die een zeer lage latency vereisen kan dit een probleem zijn, het is zaak dit type opvragingen, die voor de gegevenslaag toegang nodig hebben, te compartimenteren, in omvang (#queries en aantal personen die dit kunnen) klein en beheerst te houden en te loggen

[80] Zie Privacy by design & default, architectuurkaders

[81] Technisch vaak geïmplementeerd middels een API, verantwoorden en vastleggen van deze 'gegevenslevering' kan dan middels API management.

	A	11	De bronlaag bestaat altijd uit twee delen; de implementatielaag waar de mutaties en interacties op plaats hebben en een replica/CDC (change data capture) t.b.v. de set-en subject- gebaseerde gegevensdienst. Belangrijk; beide delen zijn integraal onderdeel van 'in productie' zijn en vallen dus onder verantwoordelijkheid van het betreffende domein.
	A	12	Bij het delen van data (system-to-system) mag de ontvangende partij de data alleen consumeren en niet verder opwerken.
	A	13	Het is niet toegestaan om gegevens te verkrijgen uit bronadministraties als daar geen beheerd artefact op zit (scherm, API, Rapport). Aka; direct gegevens verkrijgen uit de bronlaag. Streven is om ongerichte opvragingen uit bronadministraties te services uit de Generieke data access laag OF uit de gegevenslaag.
	A	14	Elke vorm van gegevenslevering (of het nu een transactie/mutatie betreft of een batchbestand) wordt gevalideerd tegen het logische model aan. Validatieresultaten worden vastgelegd en feedback naar de aanleveraar wordt altijd gegeven. Dit kan een basale status zijn maar ook inhoudelijke meldingen.
	G	15	De Centrale Feitl Laag is: ▪ Gebaseerd op de aanlevermodellen van domeinen; ▪ Wordt gerealiseerd door middel van semantisch equivalente structuurtransformaties (voorkomen wordt dat er upstream allerlei waarheden worden geïmplementeerd waarmee traceerbaarheid en multirealiteiten niet meer mogelijk zijn); ▪ Altijd geïntegreerd op concepten die semantisch identiek zijn; ▪ Volledig temporaal (in tijd) gestapeld (unmutable); ▪ Nooit direct exposed naar gebruik(ers).
	G	16	De generieke data access laag: ▪ Is altijd minimaal gebaseerd op een logisch datamodel ▪ Ondersteunt multirealiteiten (verschillende uitlevermodellen maken op dezelfde feiten); ▪ Vormt een abstractielaag t.a.v. de Centrale Feit Laag; ▪ Ondersteunt het reizende-nu dataprincipe; ▪ Ondersteunt verschillende gebruikerstypen (bv. CRUD, Management Informatie en Analytics); ▪ Is de enige wijze van toegang op de Centrale Feit Laag; ▪ Is zoveel mogelijk gevirtualiseerd en alleen gematerialiseerd als er technische concerns zijn (bv. performance); ▪ Kan een grote verscheidenheid aan technische verschijningvormen bevatten (bv. API, Views, etc..).
	G	17	Ontvangers van gegevens van de Belastingdienst krijgen die gegevens bij voorkeur beschikbaar gesteld i.p.v. (fysiek) geleverd.
	G	18	Dynamische[82] depseudonimisering moet worden toegepast in de generieke data access laag, ofwel

[82] Dynamische pseudonimisering is het streven, dat wil zeggen op runtime, afhankelijk van AVG kenmerken en autoristaties wordt data gepseudonimiseerd ter beschikking gesteld. In de Privacy by Design implementatie bij DF&A bleek dat er performance problemen waren als e.e.a. dynamisch werd gemaakt er is daarom voor gekozen e.e.a. vooralsnog in batch te pseudonimiseren.

			o.b.v. de gegevensclassificatie en de authenticatie van de gebruiker wordt bepaald of persoonsgebonden gegevens gedepseudonimiseerd getoond mogen worden. Hiermee wordt zoveel mogelijk invulling gegeven aan het principe dat gebruikers van data überhaupt de fout niet in kunnen gaan aangaande het consumeren van privacy gevoelige data.
	G	19	De wijze van uitvragen bij derden is technisch en logisch gestandaardiseerd en gepubliceerd. ▪ *Ook de technische kanalen[83] van leveren zijn gestandaardiseerd en gepubliceerd;* ▪ *Ook de variëteit aan technische formaten van aanleveringen is geminimaliseerd, volgt zoveel mogelijk open standaarden[84] en is dus gestandaardiseerd.*
Data Governance (DGO) *Set-gebaseerde gegevensdienst, type 1*	G	1	Alle aanleveringen, uitleveringen en ter beschikking stellingen zijn gekoppeld aan een geadministreerde afspraak en verplichting. De openstaande afspraken en verplichtingen zijn altijd beschikbaar ter inzage voor alle partijen (binnen en buiten de Belastingdienst).
	G	2	Alle aanleveringen worden logisch (inhoudelijk) gevalideerd op het logisch model, resultaten worden altijd teruggekoppeld aan de aanleveraar.
	G	3	Alle informatieproducten welke gevraagd worden hebben een geformuleerd doel, een eigenaar, duurzaamheidlabel[85] en een risicoprofiel. De wijze van voortbrenging wordt bepaald aan de hand van deze variabelen.
	G	4	Informatieproducten met een laag duurzaamheidslabel en laag risicoprofiel mogen *opportunistisch* worden uitgevoerd, of wel; opwerking van gegevens mag deels buiten de set-gebaseerde gegevensdienst plaatsvinden. Informatieproducten met een hoog duurzaamheidslabel een hoog risicoprofiel moeten *systematisch* worden uitgevoerd, ofwel, de opwerking moet in zijn geheel plaatsvinden binnen de set-gebaseerde gegevensdienst.
	G	5	Informatieproducten die opportunistisch zijn gemaakt hebben een van de volgende twee statussen: ▪ Wanneer wordt het betreffende product verwijderd en/of gearchiveerd; ▪ Wanneer wordt het betreffende product systematisch gerealiseerd.
		6	Data t.b.v. opportunistische informatieprodukten mag alleen buiten de set-gebaseerde dienst worden opgewerkt als deze gecompartimenteerd is.
	G	7	Het gegevensaanleverende en/of vragende domein wordt o.b.v. afspraken aangesproken (gegevensleveringsovereenkomst) op de kwaliteit van haar gegevens/logica en MOETEN deze remediëren.

[83] Denk aan system-to-system koppelingen als Logius m.b.t. fysieke leveringen, een webportaal m.b.t. upload functionaliteit en een API koppeling m.b.t. het zelf ophalen van gegevens

[84] Denk aan XBRL, XML, CSV, JSON,etc..

[85] Met 'duurzaamheidlabel' wordt bedoeld de vereist quality of service van het informatieproduct

	G	8	Het gegevensaanleverende en/of vragende domeinen levert functionele kennis t.a.v. de gegevens en logica op het gebied van definitie, betekenis, classificatie, formalisatie, integriteit en validatie Zie ook [DLA.7].
	G	9	Zowel aan de uitvragende kant als aan de ter beschikking en leverende kant wordt het principe van dataminimalisatie[86] toegepast: *Dataminimalisatie houdt in dat bij het verzamelen en verwerken van persoonsgegevens niet meer gegevens mogen worden gebruikt dan nodig is om het doel waarvoor ze gebruikt zullen worden te bereiken.* Belangrijk: in bovenstaande definitie wordt specifiek gesproken over persoonsgegevens in het kader van de AVG. Echter, bij het uitvragen van gegevens van derden danwel het leveren of ter beschikking stellen van data wordt de term dataminimalisatie ruimer geïnterpreteerd. Er wordt daarbij niet meer gegevens uitgevraagd, geleverd of ter beschikking gesteld dan strikt noodzakelijk voor het geformuleerde doel.
	G	10	Elke uitvraag bij derden dan wel ter beschikking stelling (verwerking) is altijd gebaseerd op een wettelijke taak of plicht (eventueel op basis van supra- of internationaal recht), noodzakelijk en proportioneel.
	G	11	Doorleveringen van derden aan derden zijn niet toegestaan. Vanuit het principe dat gegevens uit de bronadministratie komt waar de data gecreëerd wordt is het niet toegestaan om reeds uitgevraagde gegevens van die bron – zonder verrijking met Belastingdienst context/gegevens – door te leveren naar een derde. Deze derde moet dan een verzoek doen rechtstreeks bij de organisatie waar de bronadministratie is gepositioneerd.
	G	12	Pseudonimiseren vindt zo vroeg mogelijk plaats in de datalogistieke verwerking van gegevens[50]. Rationale is dat zelfs iemand die toch – op de een of andere wijze toegang verkrijgt tot de gegevenslaag – alleen gegevens ziet waar de persoonkenmerken gepseudonimiseerd zijn.
	G	13	Despeudonimiseren vindt zo laat mogelijk plaats in de gegevensoverdracht naar de uiteindelijke afnemer[50]
	A	14	De gegevenslaag is de minimale laag waar opvragingen en opportunistische dataopwerkingen mogen starten. Opvragingen in de bronlaag mogen alleen als daar technische redenen voor zijn.
		15	Omdat er processtappen en opwerkingen plaatsvinden is de quality of service die geleverd kan worden, denk aan actualiteit, altijd direct afhankelijk van de mate waarin de aanleverende systemen hun gegevens op orde hebben.
Data Governance (DGO) *Set-gebaseerde gegevensdienst, type 2*	G	16	Bedrijfsonderdelen hebben periodiek inzicht in de voorraad van polygestructureerde bestanden en de aanwezigheid van eventuele persoonsgebonden gegevens. Het governance uitgangspunt is hier: - Weghalen/verwijderen. Of - Annoteren en archiveren[87].

[86] Zie ook Privacy by Design & Default – Architectuurkaders
[87] Referentie Architectuur documenthuishouding en archief

Data Governance *Subject-gebaseerde gegevensdienst* (DGO)		17	Gegevens in de bronlaag zijn ontkoppeld door een subtypering aan te brengen in deze laag (zie [DLA.11]). Hierdoor kunnen gegevens subject georiënteerd, aan verschillende eisen voldoen (latency van near real time tot binnen 24 uur, hoog beschikbaar, 7x24 uur, 1 x per dag) en verschillende gradaties in kwaliteit beschikbaar worden gesteld.
Data Governance (DGO)	A	18	Het wijzigen van gegevens vindt altijd plaats op de authentieke bron. *NB. Dit staat los van de wijze van implementatie, bijvoorbeeld synchroon of a-synchroon*
	A	19	Alle gegevensleveringen en ter beschikkingstellingen worden geregistreerd. Ook de gegevensleveringen die informeel van aard zijn: *Het resultaat van een opportunistische opwerking van gegevens (kwadrant III en IV), bijvoorbeeld een bestand met gegevens, wordt verstuurd naar een partij binnen of buiten de Belastingdienst. Dit is een zeer informele gegevenslevering, maar moet wel degelijk deel uitmaken van de metadata t.a.v. informatieverwerkingsstromen. Heel concreet; de capabilties van het analytics domein stellen analisten in staat om informatieproducten naar buiten te sturen. Ervan uitgaande dat dit voldoet aan informatiebeveiligingsaspecten, moet de metadata m.b.t. deze gegevenslevering worden vastgelegd; wie heeft wat wanneer op welke wijze verstuurd aan welke ontvanger.*
	A	20	Elke vorm van logica die eenduidig is toe te wijzen aan een ketenproces wordt niet door de set-gebaseerde gegevensdienst geleverd. Ofwel wordt de logica als een service ter beschikking gesteld, ofwel wordt het resultaat van de logica ter beschikking gesteld. *Voorkomen moeten worden dat de set-gebaseerde gegevensdienst logica moet achterhalen, kopiëren, in sync houden met de logica van het betreffende ketenproces.* *Dit is een van de redenen waarom bestaande gegevensdiensten zoals DSS, EDW, Datafundamenten zo groot worden en moeilijk beheersbaar zijn.*

Bijlage A: Het Data Kwadranten Model

Onderstaand is een extract uit een artikel gepubliceerd in het Tijdschrift voor Toezicht (september 2018), geschreven door Ronald Damhof, Wouter van Aerle en Frank Ouddeken. E.e.a. is geschreven vanuit het perspectief van De Nederlandsche Bank.

Om het bestuurlijk en management kader in de uitwerking van het inrichtingsvraagstuk voor een datavoorziening handvatten te geven, is het Data Kwadrantenmodel ontwikkeld[88],een begripsvormend raamwerk rondom communicatie, organisatie, governance, beleid, investeringen, architectuur en prioriteit op het datadomein. Het is een eenvoudig model dat veel, met name publieke organisaties, in hun benadering van dit vraagstuk hebben overgenomen of dat door hen wordt toegepast.

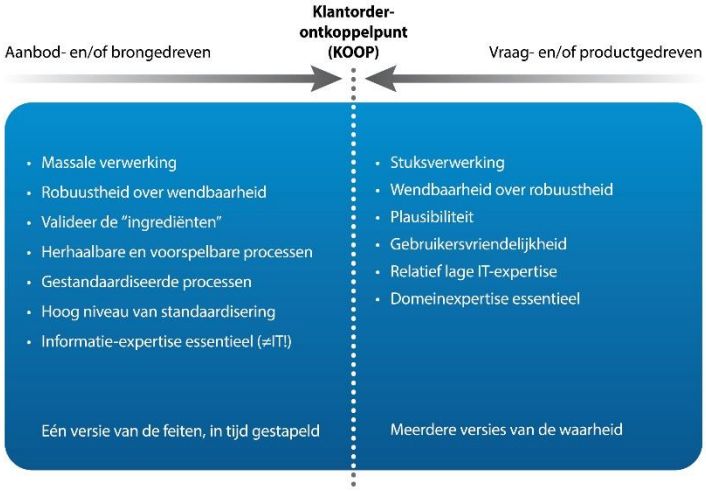
Het model is geïnspireerd door de wijze waarop ruwe grondstoffen in de fysieke wereld worden verwerkt tot eindproducten. Eenzelfde uitdaging is er in de wereld van data. Overheidsorganisaties ontvangen van verschillende partijen binnen en buiten haar organisatie steeds meer data. Van deze data (lees: ruwe grondstoffen) moeten verschillende informatieproducten worden gemaakt. Dit varieert van gebruikelijke informatieproducten als rapportages en dashboards om besluitvorming te verbeteren, tot datasets voor geavanceerde analyses, maar ook verplichte dataleveringen aan ketenpartijen in het publieke domein.

Meer ambitieus zijn informatieproducten die het resultaat zijn van algoritmes op data die een bedrijfsproces aansturen, bijvoorbeeld een verhoogd risico op een rechtspersoon waar nader onderzoek nodig is. De kern is dat data binnenkomen als ruwe grondstof en zodanig getransporteerd en verwerkt moeten worden dat de vele verschillende soorten afnemers er hun gevarieerde doelstellingen mee kunnen bereiken. In deze logistieke beschouwing van gegevensvoortbrenging zijn er twee assen die vier kwadranten opleveren. Deze assen zijn het *klantorderontkoppelpunt* en de *wijze van ontwikkeling*.

[88] Dit model is geïnspireerd op (1) The Toyota Way die haar oorsprong heeft in het werk van W. Edwards Deming, (2) het Cynefin Framework van Dave Snowden en (3) ontwikkelingen in data-architectuur van de afgelopen tien jaar met name op het gebied van feitmodellering, integratiemodellering en het ondersteunen van multi-realiteiten. ‘Maak datamanagement bespreekbaar in de hele organisatie’, in: Future Bright. A data driven reality, SAS Nederland 2015, p. 24-31, https:// issuu. com/ sasnederland/ docs/ futurebrightdatamanagement_ nl.

Klantorderontkoppelpunt

In het proces waarin ruwe grondstoffen verwerkt worden tot eindproducten is het concept klantorderontkoppelpunt (KOOP) cruciaal: tot waar dringt de vraag van de afnemer door in het verwerkingsproces? Bij het kopen van een luxe motorjacht zal er pas geproduceerd gaan worden als de order van de klant geaccordeerd is (vraaggedreven). Het KOOP ligt in dat geval bij de ruwe grondstoffen, dus bij de start van het verwerkingsproces. Bij het kopen van lucifers daarentegen gaat de afnemer naar de supermarkt, haalt ze van het schap en rekent af (aanbodgedreven). Het KOOP ligt hier bij het eindproduct. Het KOOP wordt zo onderscheiden door aanbodgedreven (push) en vraaggedreven (pull) zijde.



Kenmerken van deze twee kanten zijn onder andere:

- push is aanbodgedreven, pull is vraaggedreven;
- push is gericht op massaverwerking, pull is gericht op stukverwerking;
- push is herhaalbaar, voorspelbaar en doorloopt een uniform proces;
- push stelt betrouwbaarheid boven wendbaarheid, pull stelt wendbaarheid boven betrouwbaarheid;
- push kent een grote mate van specificatie en standaardisatie, pull heeft dat per definitie minder;
- push kent een hoge automatiseringsgraad, pull een lagere.

Figuur 14: KOOP Data Kwadranten Model

In het verwerken van data wordt bovenstaande analogie overgenomen en ontstaan er twee gebieden die fundamenteel van elkaar verschillen (zie Figuur 1). Aan de aanbodzijde (push, linkerkant) is er één versie van feiten. Vanuit het perspectief van de organisatie worden alle gegevens zoals aangeleverd door rapporteurs, technische systemen of andere instituties, beschouwd als feiten. Feiten kennen maar één versie, de versie zoals die is ontvangen, ongeacht eventuele onvolkomenheden die de data bevatten. Er is geen interpretatie nodig om deze gegevens op te slaan en te verwerken. Deze feiten zijn over de tijd heen gestapeld en worden in principe[89] niet verwijderd. Het met de data door de tijd reizen wordt daarmee mogelijk gemaakt. Allemaal niet-functionele eisen die passen in de principes van goed toezicht. Aan de rechterkant (pull, vraagzijde) zijn er meerdere versies van de waarheid. Waarheden, ook wel realiteiten genoemd, zijn

[89] Tenzij specifieke wetgeving zoals bijvoorbeeld de AVG dit verplicht.

perspectieven op feiten. Verschillende gebruikers kunnen op basis van dezelfde feiten verschillende perspectieven hebben die allemaal binnen hun specifieke context geldig en juist zijn. Waarheden zijn als het ware lenzen op de werkelijkheid. Met lenzen van verschillende sterkten en kleuren krijgen afnemers een beeld van de werkelijkheid die past binnen hun context. Dit kan bij DNB een toezichtperspectief zijn op banken, een macrostatistisch perspectief op basis van de betalingsbalans of een perspectief vanuit financiële stabiliteit.

Wijze van ontwikkelen

Een deel van de informatieproducten die gemaakt moeten worden, kent redelijk vastomlijnde eisen. Denk aan managementinformatie met betrekking tot bedrijfsvoering of het periodiek leveren van datasets aan de CBS, gemeenten, Politie en andere ketenpartners. Het maken van deze informatieproducten kent een verloop dat te vergelijken is met ordentelijke systeemontwikkeling in de informatietechnologie. We noemen dit ook wel systematisch ontwikkelen (zie Figuur 15).

Figuur 15: Ontwikkelstijl, systematisch vs opportunistisch

Binnen de DNB maar zeker ook daarbuiten is echter sprake van een niet te stoppen en potentieel waardeverhogende trend om innovatief en experimenteel om te kunnen gaan met data, vaak onder de noemer van data science. Onderzoekers, econometristen, statistici of andere kwantitatieve specialisten zoeken naar patronen, 'outliers' en inzichten die verstopt zijn in complexe en grote hoeveelheden data en die een indicatie geven waar nader onderzoek nodig is. De context waarin deze experts werken, moet een grote mate van vrijheid kennen. We noemen dit ook wel opportunistisch ontwikkelen. Het onderscheid tussen deze twee vormen van ontwikkelen is cruciaal voor het datadomein omdat de twee vormen een fundamenteel andere inrichting en operationalisatie behoeven. Hieronder benoemen we enkele onderscheidende kenmerken:

- Er is sprake van functiescheiding tussen gebruiker en ontwikkelaar in de systematische ontwikkeling. Deze is niet of beperkt aanwezig bij opportunistische ontwikkeling. Hier is degene die de data verzamelt, duidt, prepareert, schoont, analyseert, interpreteert en er soms zelfs actie op onderneemt, een en dezelfde persoon.

- Systematische ontwikkeling wordt gekenmerkt door een defensieve governance, in tegenstelling tot de opportunistische ontwikkeling waar sprake is van offensieve governance. Defensieve governance betekent een planmatige, vastomlijnde en geformaliseerde manier van ontwikkelen, implementeren en beheren van gegevensverwerkingsprocessen. Denk hierbij aan het inrichten van systeemmonitoring, wijzigingsmanagement, OTAP voortbrenging, segmentering van IT-omgevingen enzovoort. Bij een offensieve governance is dit veelal niet aanwezig en ook niet nodig. De beheersing van de gegevensverwerking is hier de volledige verantwoordelijkheid van de degene die de verwerking uitvoert.

- Systematische ontwikkeling tendeert naar een centrale inrichting binnen een organisatie, in tegenstelling tot de opportunistische wijze van ontwikkeling die een grote mate van decentralisatie (soms tot de individuele onderzoeker) kent.

- De systematische wijze van ontwikkelen vormt zich rondom de niet-functionele eisen als duurzaamheid, betrouwbaarheid, schaalbaarheid en herhaalbaarheid. De opportunistische wijze van ontwikkeling is meer gericht op eisen als wendbaarheid, snelheid en experimenteren.

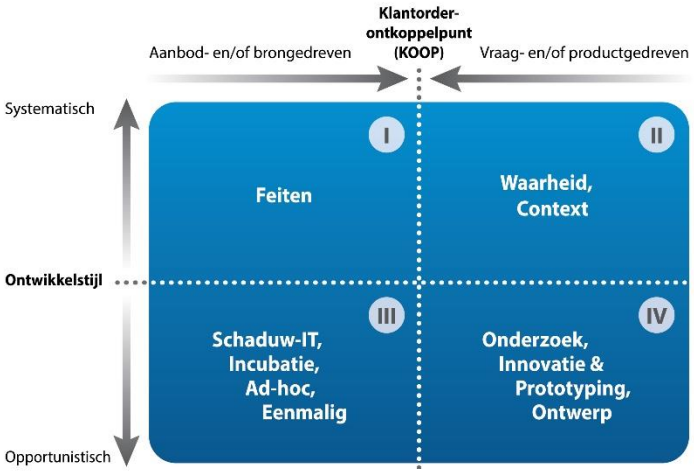
- De technische omgeving voor systematische voortbrenging is ingericht volgens standaarden die in de informatietechnologie gewoon zijn en bedoeld zijn voor langdurig, structureel en meervoudig gebruik. In de opportunistische omgeving daarentegen worden er feitelijk drie componenten aangeboden: data, computerkracht, analyse- en programmeergereedschappen. Dit soort omgevingen wordt vaak aangeduid met termen als datalab, sandbox of pilot-omgevingen[90] en kennen doorgaans een tijdelijk karakter voor de duur van een specifieke data-analyse.

[90] Zie bijvoorbeeld www. toezine. nl/ artikel/ 228/ ilt -verbetert -datagedreven -toezicht -met -datalab/ .

De kwadranten nader belicht

De twee assen gecombineerd geven een viertal kwadranten (zie Figuur 16).

Het is belangrijk te benadrukken dat de assen waardevrij zijn. Kwadrant I is niet beter, mooier of belangrijker dan kwadrant IV. Hoe iedere organisatie de kwadranten invult, is contextafhankelijk. Voor toezichthouders en formele datasets die gebruikt worden voor de risicodetectie op inkomensheffing is het van



Figuur 16: Data Kwadranten Model (Damhof)

belang dat de gegevensverwerking transparant en auditeerbaar is (kwadrant I). Voor onderzoekers is het opportuun dat ze eigen data (kwadrant III) kunnen combineren met systematisch voortgebrachte data (kwadrant I) op zodanige wijze dat ze met gespecialiseerde analysegereedschappen (kwadrant IV) kunnen experimenteren.

Kwadrant I: Feiten

In kwadrant I staan de gegevens zoals ze bij aanlevering zijn aangeboden. Er wordt gestreefd naar een minimale interpretatie van gegevens en maximale precisie en ondubbelzinnigheid. De automatiseringsgraad en de mate van beheersbaarheid zijn in dit kwadrant maximaal. Bij het doorvoeren van wijzigingen in de gegevensverwerking zal beheersbaarheid hoger worden gewaardeerd dan wendbaarheid. In de context van een toezichthouder DNB worden in kwadrant I hoge eisen gesteld aan de verwerking, met name ten aanzien van traceerbaarheid, beheersbaarheid, schaalbaarheid en betrouwbaarheid.

Kwadrant II: Context

In dit kwadrant worden informatieproducten gemaakt en in beheer genomen die structurele behoeften van de gebruiker vertegenwoordigen. Deze informatieproducten komen tot stand door een systematisch productieproces waarbij gebruik wordt gemaakt van de feiten in kwadrant I en waarbij vooraf duidelijk is verwoord wat de kwaliteitseisen zijn en op welke wijze de informatieproducten in de bedrijfsprocessen van waarde zijn. In dit kwadrant kunnen informatieproducten worden gerealiseerd die op dezelfde feiten zijn gebaseerd maar als gevolg van een verschillend perspectief op die feiten van elkaar verschillen of elkaar zelfs tegenspreken (ondersteuning multirealiteiten).

Kwadrant III: Shadow-IT, ad hoc, eenmalig

Elke organisatie heeft een kwadrant III. Denk aan alle Excelbestanden, MS Access databases en andere lokale gegevensdragers. Kenmerk van dit soort omgevingen is dat de enige persoon die de data begrijpt, vaak ook diegene is die de data er heeft neergezet. Het is ook het kwadrant waar het mogelijk moet zijn om data neer te zetten die een eenmalig of onvoorspelbaar karakter hebben. Het is hierdoor ook niet nodig om deze data in kwadrant I te plaatsen. Belangrijk is verder dat de gegevens in kwadrant III onder beperkt georganiseerd beheer staan en dat de betreffende datakwaliteit hoogstens bekend is bij degene die de data er heeft neergezet. Let wel: ook in dit kwadrant staan de data wel degelijk onder governance. Alleen is het de gebruiker zelf die de niet-functionele governance-eisen moet handhaven.

Kwadrant IV: Research, innovatie, prototyping

Kwadrant IV staat in het teken van het adagium ‘experimenteer, leer en verbeter’. Het is het kwadrant waar medewerkers vragen om (bij voorkeur goede) data, computerkracht en geavanceerde software om data-analyses uit te voeren. Het is het kwadrant waar veel vrijheid wordt geboden en veel wordt gevraagd van de verantwoordelijkheid van de betreffende medewerkers. Kwadrant IV wordt voor verschillende toepassingen gebruikt. Veel voorkomend is prototyping: het iteratief ontwikkelen van een basisversie van een informatieproduct. Prototyping is bij uitstek geschikt wanneer een gebruiker een globaal idee heeft van zijn informatiebehoeften en deze graag in de vorm van een voorbeeld uitgewerkt wil zien. Samen met de gebruiker wordt dan vaak in een aantal iteraties het gewenste product ontwikkeld. Ad-hoc analyses zijn een andere toepassingsvorm. In feite zijn dit alle mogelijke soorten niet-standaardanalyses op reeds aanwezige data. Research is een andere toepassingsvorm, zoals onderzoek naar nieuwe risicomodellen voor kredieten, analytische voorspelmodellen voor solvabiliteit bij verzekeraars of algoritmes die outliers vroegtijdig identificeren. Daarnaast is kwadrant IV de aangewezen omgeving voor innovatie, zoals het uitproberen van nieuwe analytische technieken als machine learning. Ten slotte is kwadrant IV – vaak in combinatie met kwadrant III – de Haarlemmerolie van een datavoorziening. Voor situaties waarin er direct gereageerd moet worden zoals bij een calamiteit of

andere niet van tevoren bedachte vormen van gebruik, biedt kwadrant IV de ruimte om naar eigen inzicht data te verwerken en analyses uit te voeren.

Van visie naar inrichting

Het kwadrantenmodel wordt door vele organisaties gebruikt als referentiekader bij het doorvoeren van structurele vernieuwingen in de wijze waarop gegevens worden verwerkt voor toezicht en de macrostatistiek. Hierbij heeft het model niet alleen zijn toegevoegde waarde bewezen bij het (opnieuw) ordenen en inrichten van de (fysieke) gegevensverwerkingen maar ook bij het bepalen van de hiervoor gewenste organisatorische inrichting, besturing, benodigde kennis en competenties en noodzakelijke technologieën. Deze invalshoeken komen hieronder achtereenvolgens aan de orde.

Keuzes in voortbrengingspatronen

Eerder is gemotiveerd dat een toezichthouder of elke andere overheidsorganisatie te maken heeft met diverse niet-functionele eisen als traceerbaarheid, vertrouwelijkheid, toegankelijkheid en betrouwbaarheid. Het kwadrantenmodel maakt duidelijk hoe de governance van deze eisen verloopt. Aan de hand van het model kunnen nu keuzes gemaakt worden welke patronen van gegevensvoortbrenging wel en niet zijn toegestaan.

- Het systematische patroon: van KI naar KII. Dit patroon wordt gevolgd voor gegevensvoortbrengingen die structureel en periodiek moeten worden uitgevoerd en waarbij de niet-functionele eisen (geautomatiseerd) moeten worden gewaarborgd. Voorbeelden zijn periodieke doorleveringen van CRDIV en Solvency II-rapportages naar de ECB resp. EIOPA en dashboards voor het toezicht op de verzekeringssector.

- Het opportunistische patroon: KIII naar KIV. Hier speelt de gegevensvoortbrenging zich vrijwel volledig af in de persoonlijke werkomgeving van een gebruiker of onderzoeksgroep. Voor een specifiek doel worden gegevenssets (handmatig) ontsloten en vervolgens gebruikt voor analyse. Een voorbeeld is een onderzoek van DNB naar de procycliciteit van initiële margevereisten van renteswaps (IRS), waarbij gebruik wordt gemaakt van informatie over de dagelijkse posities ontleend aan verschillende transactieregisters en geavanceerde analytische modellen.

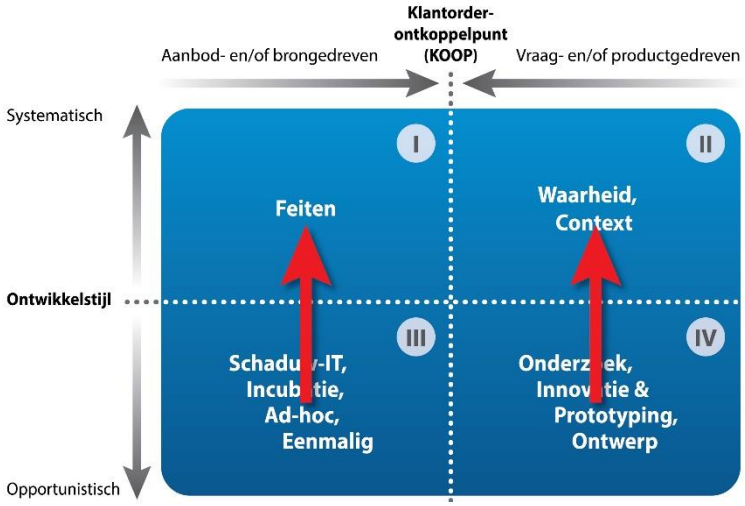
- Het wendbare patroon: KI naar KIV: in dit patroon worden systematisch ontvangen en opgeslagen data vervolgens voor analytische doeleinden gebruikt. Zo worden de CRDIV-data nadat ze zijn ontvangen gevalideerd en opgeslagen, daarna ook beschikbaar gesteld voor analisten in KIV die ieder vanuit hun eigen werkomgeving ad-hoc analyses op deze dataset willen uitvoeren. In de praktijk is dit een veel gebruikt patroon dat door gebruikers en analisten hoog gewaardeerd wordt. Ze hoeven immers niet zelf voor de (periodieke) ontsluiting

en validatie van de data te zorgen maar krijgen wel vrijheidsgraden in het zelf doorzoeken en analyseren van dit soort datasets.

Vanuit het oogpunt van verantwoording en risicobeheersing is het daarentegen ongewenst dat opportunistisch voortgebrachte data (kwadrant III) worden gebruikt in formele informatieproducten (kwadrant II). Indien deze behoefte toch bestaat, is het nodig de gegevensontsluiting in kwadrant III te promoveren naar kwadrant I, ofwel een degradatie van het informatieproduct in kwadrant II naar kwadrant IV (waarmee het onder verantwoordelijkheid komt van de data-analist of gebruiker die initieel het informatieproduct heeft ontwikkeld).

De ultieme queeste: van kwadrant III/IV naar kwadrant I/II

De grootste uitdaging bij het realiseren van datagedreven toepassingen is het structureel toepassen van bruikbare algoritmes, geslaagde experimenten en prototypen in reguliere toezichtprocessen. In termen van het kwadrantenmodel gaat het hierbij om de beweging van kwadrant III/IV naar kwadrant I/II (zie Figuur 17).



Figuur 17: Operationalisatie van datascience

Van een algoritme dat 'werkt op mijn laptop' moet nu een permanent werkend en periodiek geautomatiseerd voortgebracht informatieproduct gemaakt worden dat schaalbaar is naar een grote(re) groep gebruikers en 'by design' voldoet aan de gestelde niet-functionele eisen. Uit ervaringen in de praktijk blijkt dat hier een forse uitdaging ligt die voortkomt uit een brede kloof tussen kwadrant I/II en kwadrant III/IV. Die kloof wordt gevormd door verschillen in de manier van werken (systematisch versus opportunistisch), profiel van betrokken medewerkers (de 'suits' versus de 'hoodies'), gebruikte technologieën en organisatorische positionering (centraal versus decentraal). Het is van belang dat leidinggevenden erkennen dat alle kwadranten van cruciaal belang zijn en dat de randvoorwaarden worden gecreëerd voor optimale samenwerking en afstemming. Het ontwikkelen van een dergelijke *organisatorische competentie* blijkt in de praktijk een van de grootste uitdagingen in het werken met data te zijn en moet op vele vlakken gerealiseerd worden: cultuur, mens, proces en technologie. Dit moet gezien worden als een forse culturele transformatie waarbij nieuwe vormen van samenwerking tussen het datadomein en het toezichtdomein gevonden en opgezet moeten worden.

Bijlage B: Logische Modellering

Een logisch datamodel (LDM) is de universele formele taal waarmee systeemartefacten met data communiceren.

- Wiskundige formele complete beschrijving van een informatieschema.
- Het logisch datamodel is een relationeel datamodel, gebaseerd op Codd (~ 1970) – First Order Logic en Set theory, aangevuld door Sjir Nijssen in 1980. De notatie/specificatie is die van Chen (~1976), Entity Relationship Diagramming.
- Formaliseert de verbanden tussen begrippen, waardoor
  - het fysieke domein geheel onafhankelijk is van het conceptuele domein;
  - data-onafhankelijkheid bereikt wordt: in het fysieke domein is alles toegestaan, zolang de data conform het logische datamodel gereproduceerd wordt.
- Het LDM is de standaard voor de voortbrenging van software en rapportages.
- Het LDM is technologisch agnostisch, er zijn geen technisch specifieke belangen vertegenwoordigd, zoals performance en partitionering.
- Het LDM is de standaard voor het modelleren van gegevens. Het LDM wordt altijd gemaakt.
- In geval van een leveringssituatie is het LDM onderdeel van de GLO en dus samen met de andere onderdelen van de GLO definieert het de levering.
- De bovenliggende niveaus van representatie zijn leidend voor de representatie in het LDM. Op het logische niveau heerst keuzevrijheid over een bepaald issue zolang een hoger niveau zich niet over dit issue uitspreekt.
- Concepten in het LDM zijn binnen hun namespace altijd uniek, bij een collision wordt overleg gepleegd met de namespace belanghebbenden.
- Het LDM vormt een gegevensverzameling voor kolommen, als basis voor kolom gebaseerde toegang tot de data.
- Algebra: Tutorial D en het type-systeem uit The Third Manifesto[91].

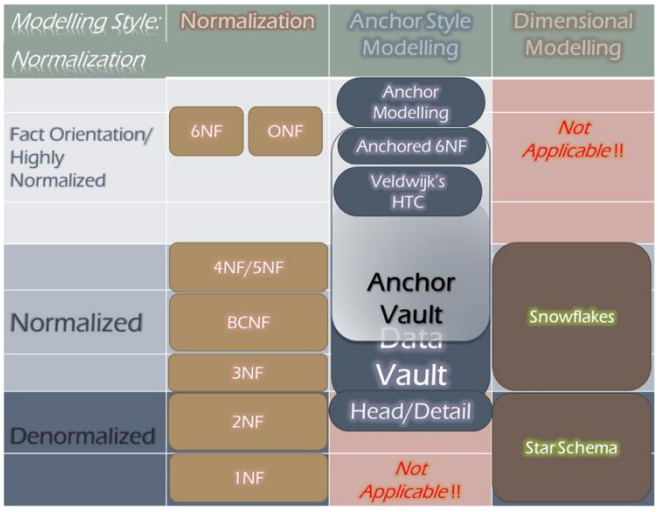
Belangen:

- Primaire belangen: wiskundige en formele representatie, manipulatie, beperking, afleiding, typing van data.
- Secundaire belangen: classificatie, integratie.

[91] Zie http://www.thethirdmanifesto.com/

Bijlage C: Anchor Style Implementatiemodellering

De verantwoording bij de keuze van modelleer strategie in de Centrale Feit Laag (CFL) begint met het toelichten van het zogenaamde 'Anchor Style' modelleren. Dit is een vorm van modelleren die gekenmerkt wordt door het scheiden van sleutels (keys) en context (attributen). Deze vorm van modelleren moet gescheiden worden van dimensioneel of genormaliseerde vormen van modelleren (zie ook Figuur 18).



Figuur 18: Ge-anchoriseerde modellering gepositioneerd

Overwegingen

Binnen de 'Anchor Style' modelleer methoden kan een onderscheid worden gemaakt tussen een aantal methoden die zich o.a. (met name) hebben bewezen in Enterprise Data Warehouse omgevingen; Data Vault van Dan Linstedt[92], generieke temporale modelleer stijlen zoals Veldwijk's HTC benadering en de Anchor Modelleer stijl[93] die met name door Lars Ronnback gepopulariseerd is. Wat al deze stijlen gemeen hebben is dat ze hun oorsprong kennen in feiten gebaseerde vormen van modellering zoals NIAM94/ORM95[94] en FCO-IM96[95]. Een stevige historische en fundamentele basis waarbij DNB een zeer duurzame (modelleer) toekomst in gaat.

Welke modelleervorm binnen de Anchor Stijl moet worden gebruikt is uiteengezet in onderstaande, daarbij zijn de volgende eisen zijn van belang in de keuze:

- Scheiding van feiten en context;
- Correctheid, consistentie, compleetheid en het vermijden van ambiguïteit;
- Herhaalbaarheid en vaste model-transformatie patronen;
- Transparantie in de wijze van modelleren;

[92] Kijk voor meer informatie op http://danlinstedt.com of 'Modeling the Agile Data Warehouse with Data Vault'- Hans Hultgren (2012)
[93] Kijk voor meer informatie op http://www.anchormodeling.com
[94] Kijk voor meer informatie op http://nl.wikipedia.org/wiki/NIAM
[95] Met name door Terry Halpin beschreven in 'information modeling and relational databases, second edition'
[96] Kijk voor meer informatie op http://www.fco-im.nl

- Lokale compartimentering van data-model veranderingen;
- Duurzaamheid wordt met name bepaald door vaste leerbare modelleer patronen;
- De modellering-vorm moet tijd ondersteunen;
- Het moet mogelijk zijn om generalisatie/specialisatie of subtyping/supertyping te ondersteunen.

Voor het modelleren van de CFL moet gekozen worden voor een 'Anchor Stijl' wijze van modelleren.

a) Zoals ook onderliggend aan het Data Kwadranten Model, is de *separatie tussen feiten en context* een cruciale. Anchor Style modelleren maakt een duidelijk onderscheid tussen enerzijds entiteiten en anchors[97] die niet veranderen in tijd en anderzijds contextuele informatie die wel verandert in tijd en dus (traceerbaarheid, auditeerbaarheid) gevolgd moeten worden.

b) *Correctheid, consistentie, compleetheid en het vermijden van ambiguïteit* zijn cruciale non-functionals. en de wijze van modelleren bepaalt deze in grote mate. Alle 'Anchor Style' modelleer methoden ondersteunen deze non-functionals, met name door het model consistent en transparant te houden.

c) *Herhaalbaarheid en vaste transformatie patronen* zijn belangrijk om te komen tot een grote mate van *standaardisering en automatisering*.

d) *Transparantie* in gegevens (model), ofwel 'All of the data, All of the time', wordt gehaald als er strikt wordt gemodelleerd.

e) *Flexibiliteit, lokale compartimentering* van data-model veranderingen is belangrijk. Data modellen dienen zoveel als mogelijk incrementeel bijgezet te kunnen worden in plaats van dat bestaande structuren veranderen (wat zware consequenties heeft voor bestaande laad-structuren en de steeds complexere regressie testen).

f) In termen van *duurzaamheid* is het van belang dat de wijze van modelleren leerbaar is en zoveel als mogelijk gestandaardiseerd kan worden. Bovendien moet de wijze van modelleren controleerbaar door derden en/of peers. Goed geformuleerde regels voor modelleren horen daarbij.

g) De modelleer-strategie moet alle vormen van tijds-modellering kunnen ondersteunen, zijnde uni-temporaliteit (denk hierbij aan transactietijd), bi-temporaliteit (transactietijd en geldigheidstijd) en eventueel tri-temporaliteit.

h) Het moet mogelijk zijn generalisatie/specialisatie, of subtyping/supertyping, te ondersteunen. Dit heeft de volgende argumenten;

    a. een beter begrip van het onderliggende conceptuele model;

    b. context attributen van een anchor behoren altijd bij alle instantiaties van de anchor. Er is geen noodzaak voor de 'N/A' waarde;

    c. de mogelijkheid om subtype constraints toe te voegen waarmee de data kwaliteit wordt verhoogd;

[97] Het verschil tussen een entiteit en anchor in dit verband wordt nihil beschouwd.

d. Minder business rules nodig 'downstream' (bv. Kwadrant II)

i) De modelleerstrategie moet een strategie zijn waarbij nieuwe of veranderende data objecten een uitbreiding betekenen op het datamodel in plaats van een verandering van bestaande data modellen. Ofwel; veranderingen in data objecten moeten een zo lokaal mogelijke impact hebben op de totale voortbrenging. Met name de regressie testen in dergelijke omgevingen worden daarmee in control gehouden.

Binnen de 'Anchor Style' modelleer methoden is gekozen om een vorm van modelleren te gebruiken die een combinatie is van een aantal methoden; Data Vault van Dan Linstedt[98], generieke temporale modelleer stijlen zoals Veldwijk's HTC benadering en de Anchor Modeling style[99] die met name door Lars Ronnback gepopulariseerd is.

Wat al deze stijlen gemeen hebben is dat ze hun oorsprong kennen in feiten gebaseerde vormen van modellering zoals NIAM[100]/ORM[101] en FCO-IM[102]. Een stevige historische en fundamentele basis waarbij een zeer duurzame (modelleer) toekomst in wordt gegaan.

[98] Kijk voor meer informatie op http://danlinstedt.com of 'Modeling the Agile Data Warehouse with Data Vault'- Hans Hultgren (2012)

[99] Kijk voor meer informatie op http://www.anchormodeling.com

[100] Kijk voor meer informatie op http://nl.wikipedia.org/wiki/NIAM

[101] Met name door Terry Halpin beschreven in 'information modeling and relational databases, second edition'

[102] Kijk voor meer informatie op http://www.fco-im.nl

Bijlage D: Gegevensleveringsovereenkomsten[103]

Een GLO heeft drie doelstellingen:

a) *Data Governance-instrument:* beide partijen committeren zich aan de voorwaarden en eisen die zijn gesteld aan de gegevens. Ook AVG gegevenskenmerken en AVG verwerkingskenmerken[104] zijn onderdeel van de GLO. Ander aspecten die van belang zijn; periodiciteit, bestandsformaat, data- en domeintypes, integriteit, etc..

   De GLO moet worden gezien als een vorm van een Service Level Agreement die twee partijen overeenkomen en waarover door de uitvoerende partij gerapporteerd moet worden richting afnemende partij.

   In die rol als SLA moet er dus ook gerapporteerd kunnen worden naar partijen aangaande in hoeverre partijen de gestelde contractvoorwaarden ook halen. Denk bv. aan leverbetrouwbaarheid, validatiebetrouwbaarheid, etc..

b) *Data Design-instrument:* datamodelleurs zijn in staat om op basis van de GLO de verschillende fysieke modellen en mappingen te realiseren.

c) *Data Processing-instrument:* de GLO moet de gegevensverwerking voorwaardelijk aansturen. Ofwel; als er geen geldige overeengekomen GLO is dan bestaat er ook geen mogelijkheid om gegevens geautomatiseerd te verwerken.

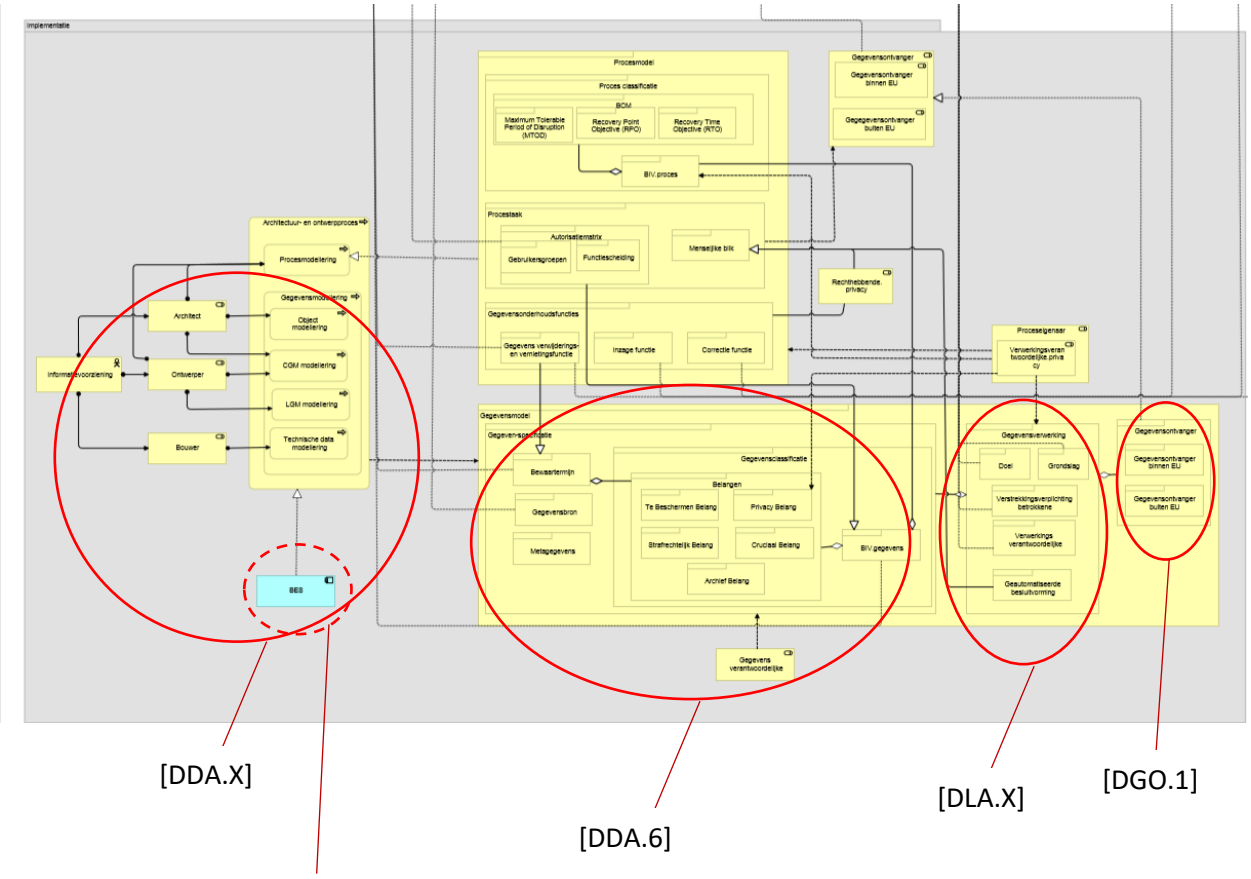
uitbreiden

[103] Een voorbeeld van een GLO (buiten de Belastingdienst) is te vinden op https://www.dnb.nl/binaries/AnaCredit%20Gegevensleveringsovereenkomst%20v1.3.1_tcm46-381589.pdf

[104] Zie ook Privacy by Design & Default, architectuurkaders

Bijlage E: Referentiearchitectuur Informatiebeveiliging

De referentiearchitectuur Informatiebeveiliging het volgende implementatiemodel gebruikt aangaande modellering en de vastlegging van allerlei kenmerken die een relatie hebben met classificaties, bewaartermijnen, etc..



De gereedschappen welke worden gekozen voor ondersteuning van het ontwerp, beheer, prubliceren en deployen van informatie- en datamodellen wordt vastgesteld in de Domein Architectuur Gegevens.

NB. In bovenstaand figuur wordt gesproken over ‘Object modellering’ en ‘CGM modellering’. In de Data Definitie Architectuur zoals beschreven in dit document, vervallen deze termen.

De (design) gegevensartefacten die relevant zijn t.a.v. bovenstaand figuur zijn:

- De *Gegevensleveringsovereenkomst* t.a.v. de (AVG) verwerkingskenmerken;
- Het *logische mode*l t.a.v. de AVG kenmerken en ander genoemde kenmerken;

- Het nog te realiseren *metamodel*[105] wat bovenstaande kenmerken vastlegt, onderhoudbaar en vindbaar maakt.

Tenslotte is van belang een aantal raakvlakken met de Referentie Architectuur Informatiebeveiliging te beschrijven. Als de kaders, zoals beschreven in dit document worden opgevolgd dan:

- is het mogelijk een up-to-date *verwerkingsregister* te realiseren[DLA.x][[DGO.1] DGO.6];
- wordt opvolging gegeven aan het *dataminimalisatie* principe [DGO.9] [DGO.10];
- zijn gegevens m.b.t. *AVG en labeling van belangen* inzichtelijk, onderhouden en opvraagbaar [DDA.6];
- is het uitvoeren van *pseudonimisering* realiseerbaar. Voor pseudonimisering is met name het werk wat bij DF&A is uitgevoerd cruciaal en leidinggevend[50] [DLA.5][DLA.7][DLA.18][DLA.13][DGO.12][DGO.13].

[105] Er zal in dit kader zoveel mogelijk hergebruikt worden van het door DF&A gemaakte metadatamodel t.a.v Privacy by Design 1.01

Bijlage F: Mapping vorige versie Referentiearchitectuur Gegevens[106]

RA Gegevens – Regels	Ref. Arch. GGHH
1. Leestoegang van gegevens uit bronsystemen voor andere doeleinden dan het operationele gebruik in het transactieproces , zoals bijvoorbeeld interactie , toezicht of analytics gebeurt altijd via 1 van de voorzieningen van het datareservoir. Op deze manier is altijd duidelijk bij de afnemer welke kwaliteit gegevens geboden wordt, kunnen de gegevens ingericht worden op de manier die aansluit bij het gebruik van de afnemer en wordt gereguleerd wie afnemer is van de gegevens.	[DLA.7] [DLA.11] [DLA.13] [DGO.14]
2. Updaten van gegevens in de bronsystemen vanuit gebieden interactie, toezicht en analytics gebeurt altijd via services die beschikbaar zijn gesteld door het bronsysteem. Op deze manier kan eenduidig beheer gewaarborgd worden.	[DGO.18]
3. Gegevens komen in het shared operational domain om de quality of service eisen van de gebruikers van de gegevens te ontkoppelen van de verwerking van de gegevens in de transactie verwerkende systemen. Hierdoor kunnen gegevens subject georiënteerd, aan verschillende eisen voldoen (latency van near real time tot binnen 24 uur, hoog beschikbaar, 7x24 uur, 1 x per dag) en verschillende gradaties in kwaliteit beschikbaar worden gesteld.	[DGO.17]
4. Uitwisseling van operationele gegevens binnen een procesketen tussen transactie verwerkende systemen verloopt niet via het datareservoir (denk aan vorderingen tussen heffings – en inningssystemen) maar gebeurt rechtstreeks. Het gaat in dit geval niet om shared operational gegevens maar gegevens t.b.v. een specifiek operationeel proces dat vraagt om de altijd actuele bron. Het gaat hierbij om het uitvoeren van functies met een hoge transactiesnelheid van crud operaties en een (near) real time latency. Hiervoor zijn verschillende implementatiemogelijkheden (events, bestandsuitwisseling met legacy, data services op de bron).	[DLA.9] [DLA.10]
5. In het geval van nieuwe koppelingen wordt voor het gebruik van masterdata in de transactieprocessen gebruik gemaakt van MIH. Gebruik van de bestaande systemen gebeurt alleen wanneer MIH de benodigde gegevens bevat of het gewenste patroon (bv bulk) nog niet ondersteund.	[DDA.12]
6. Gegevens in de bronsystemen worden default beschikbaar gesteld t.b.v. andere doeleinden (d.m.v. de gegevens integratielaag). Dit stelt eisen aan de bronsystemen zoals bijvoorbeeld de manier waarop uitbreiding aan het datamodel wordt uitgevoerd (zie ook aansluitvoorwaarden MIH).	[DLA.2] [DLA.7] [DLA.11]
7. Wanneer gegevens in het datareservoir worden opgenomen wordt de catalogus (business glossary) gevuld zodat betekenis , context en kwaliteit van de gegevens helder zijn.	[DDA.x]
8. In de catalogus worden een aantal basisgegevenssets gedefinieerd. Dit is een geconsolideerde set van gegevensdefinities rondom een subject of object. Bijvoorbeeld: persoon, onderneming, inkomen, huishouden, en	[DDA.x]

[106] Verkenning informatiepositie Klantbeeld Gegevenshuishouding, versie 12/2016, vastgesteld in IV-regieteam, ook wel referentiearchitectuur gegevens genoemd

	roerend goed. Ze zijn bedoeld voor een gestandaardiseerde manier van gegevensverstrekking. De basisgegevenssets kunnen worden geïmplementeerd zowel in het analytical en operational domein maar wel met een ander gebruiksdoel en andere kwaliteitseisen.	
9.	De basis voor de opbouw van de gegevenshuishouding ligt bij de bronsystemen. Kwaliteitsissues moeten dáár opgelost worden	[DDA.x] [DGO.20]
10.	De voorzieningen binnen de gegevenshuishouding bieden een aantal standaard services (gericht op veelvuldig gebruik, meest voorkomende vragen). Om de services af te stemmen op het gebruik van de afnemer kan de afnemer bovenop de service een API realiseren. Deze is ook in beheer van de afnemer en is dus ook geen onderdeel van de gegevenshuishouding. In principe zijn deze API's dus specifiek. Deze gelaagdheid maakt het mogelijk om de voortbrenging van systemen die om een meer agile aanpak vragen onafhankelijk te maken van de meer traditionele systemen. Als interne services maar beschikbaar zijn kunnen deze onafhankelijk hiervan op maat gemaakt worden via API's voor andere systemen .	[DLA.x] [DDA.16]

Bijlage G: Mapping Beleidsvisie Integraal datamanagement[107]

Integraal datamanagement - principes	Ref.arch GGHH
1. De Belastingdienst werkt met gegevens over burgers en bedrijven. Bescherming van die gegevens tegen ongeoorloofd gebruik of tegen verlies heeft hoge prioriteit bij elk aspect van het eigen werk en in de samenwerking met anderen. Zo is en blijft de Belastingdienst een betrouwbare organisatie.	RA Informatie-beveiliging [DDA.6] [DLA.5] [DLA.13] [DLA.18] [DGO.6] [DGO.9] [DGO.12] [DGO.13] [DGO.19]
2. De Belastingdienst verwerkt gegevens om zijn taken goed uit te voeren: heffen, uitkeren, innen, goederentoezicht en opsporing. Verwerking en beheer van gegevens zijn altijd gebaseerd op een wettelijke taak of plicht (eventueel op basis van supra- of internationaal recht), noodzakelijk en proportioneel. Gegevens worden niet langer bewaard en niet eerder vernietigd dan wettelijk is toegestaan. De bewaar- en vernietigingstermijnen zijn voor alle gegevens bepaald en bekend.	[DDA.6] [DGO.9] [DGO.10]
3. De Belastingdienst heeft een open oog voor maatschappelijke en technologische ontwikkelingen die gegevensverwerking raken. Hij gaat intern en extern de dialoog aan over kansen die gebruik van gegevens biedt en over de dilemma's die daarbij kunnen ontstaan. Hij agendeert kwesties waarin spanning ontstaat tussen wat kan en mag bij de wetgever.	N/A Meer op beleid
4. De Belastingdienst is transparant over en aanspreekbaar op de gegevens die hij verwerkt. Hij streeft ernaar burgers en bedrijven zo veel mogelijk actief inzicht te geven in hun gegevens en hen regie te geven op eigen gegevens. Gegevens worden, binnen wettelijke grenzen, als open data beschikbaar geteld.	*Kapstok van deze referentiearchitectuur* Alle kaders van dit document
5. De Belastingdienst stelt gegevens alleen beschikbaar aan andere (overheids)organisaties als wetgeving dat toestaat of daartoe verplicht, en als die organisaties voldoen aan voor hen geldende wettelijke eisen voor verwerking van die gegevens. Inzage in de bron heeft daarbij de voorkeur boven leveren uit de bron. Afspraken daarover worden vastgelegd en bekend gemaakt.	[DLA.7] [DLA.17] [DGO.1] [DGO.10] [DGO.19]
6. Bij de Belastingdienst zijn de verantwoordelijkheden, bevoegdheden en taken bij gegevensverwerking bepaald en bekend. Beslissingen over verwerking van gegevens zijn traceerbaar en kunnen worden verantwoord.	Eerste deel is beleidsmatig en organisatorisch van aard.

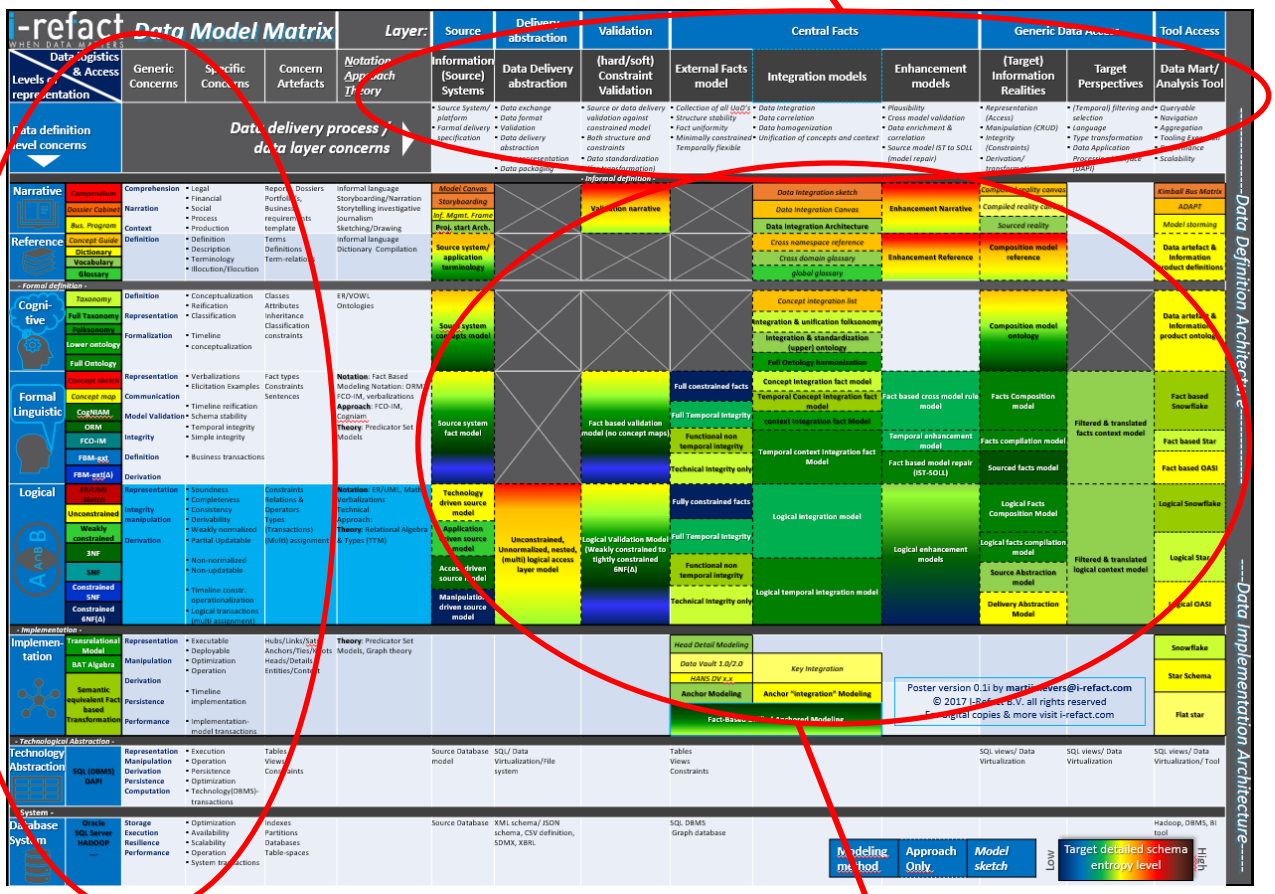
[107] Beleidsvisie Integraal datamanagement, vastgesteld in DT BD 20-11-2017

	[DLA.7] tav vastleggen besluiten rondom verwerking [DDA.6]
7. Medewerkers van de Belastingdienst (interne en externe) gaan op integere wijze om met gegevens: zij zijn zich bewust van hun verantwoordelijkheid bij het omgaan met gegevens van burgers en bedrijven en handelen daar ook naar. De Belastingdienst zorgt dat zij hierop goed zijn toegerust en waar nodig worden aangesproken.	Beleid, gedrag en cultuur aspect [DLA.5] [DLA.13] [DLA.18] heeft sterke raakvlakken [DGO.9] [DGO.12] [DGO.13]
8. De Belastingdienst streeft naar de hoogst mogelijke kwaliteit van gegevens. Belangrijke aspecten daarbij zijn actualiteit, juistheid en volledigheid. Hij waarborgt een eenduidige betekenis van de gegevens, zowel bij intern gebruik als bij het delen met burgers, bedrijven en andere organisaties.	Alle kaders van dit document
9. De Belastingdienst treft op elk punt in het proces van gegevensverwerking aantoonbaar technische en organisatorische maatregelen om de privacy te verzekeren en de kans op lekken en misbruik van gegevens zo klein mogelijk te maken.	Zie ook RA Informatiebeveiliging [DDA.6] [DLA.5] [DLA.7] [DLA.13] [DLA.18] [DGO.6] [DGO.9] [DGO.12] [DGO.13] [DGO.19]
10. De Belastingdienst maakt bij het uitwisselen van gegevens aan andere organisaties gebruik van generieke voorzieningen en van overheidsbrede en internationale standaarden die de principes van gegevensbescherming verzekeren.	[DLA.19] [DGO.9] [DGO.10] [DGO.19]

Bijlage H:Niveaus van representatie, modellering en concerns

Onderstaande concern matrix is gebruikt om zowel verticale (datadefinitie) als horizontale (datalogistieke) niveaus van representatie en concerns te onderscheiden. Deze matrix is vooral bedoeld om ‘geloofsdiscussies’ rondom modelleer-methodieken zo feitelijk en fundamenteel mogelijk te voeren over de as van het niveau van representatie en de daarbij behorende concerns.

Horizontale niveaus van representatie, aka de datalogistieke architectuur



Verticale niveaus van representatie, aka de datadefinitie architectuur

Modelleer methodieken behorende bij de combinatie van horizontale en verticale niveaus van representatie

Bijlage I: Traceerbaarheid en auditeerbaarheid

Vereisen voor het integraal bijhouden van alle operationele metadata rond dataverwerking, zowel qua data definitie, architectuur alsook datalogistiek en toegang.

1. Koppeling data/metadata op het ***laagste*** granulariteit zodanig dataverwerking 100% traceerbaar wordt (technisch: implementatie van een zogenaamd audit-id);

2. Volledige temporele non-destructieve verwerking van data en metadata;

3. Onafhankelijke tijdsconsistente toegang tot data en metadata, zowel transactioneel (bv. het doen van correcties) als batch;

4. Het ondersteunen van correcte datamodellen, ***elementaire*** structuurtransformaties en rule transformaties (anchor-style modellen, logische modellen);

5. Het ondersteunen van meerdere gecorreleerde model realiteiten.

6. Reizend nu concept

1= model-lineage en traceerbaarheid (model en mapping inzichtelijk)
2= full logging (operationele logistiek inzichtelijk)
1+2+3= auditeerbaar
1+2+3+5=volledige (110%) auditeerbaar
4+6= temporele multirealiteit
1 t/m 6= volledige auditeerbare temporele multirealiteit=110% Transparantie

Bijlage J: Data Definitie Architectuur, voorbeelden