Aan de slag met datakwaliteit

Een praktische handreiking

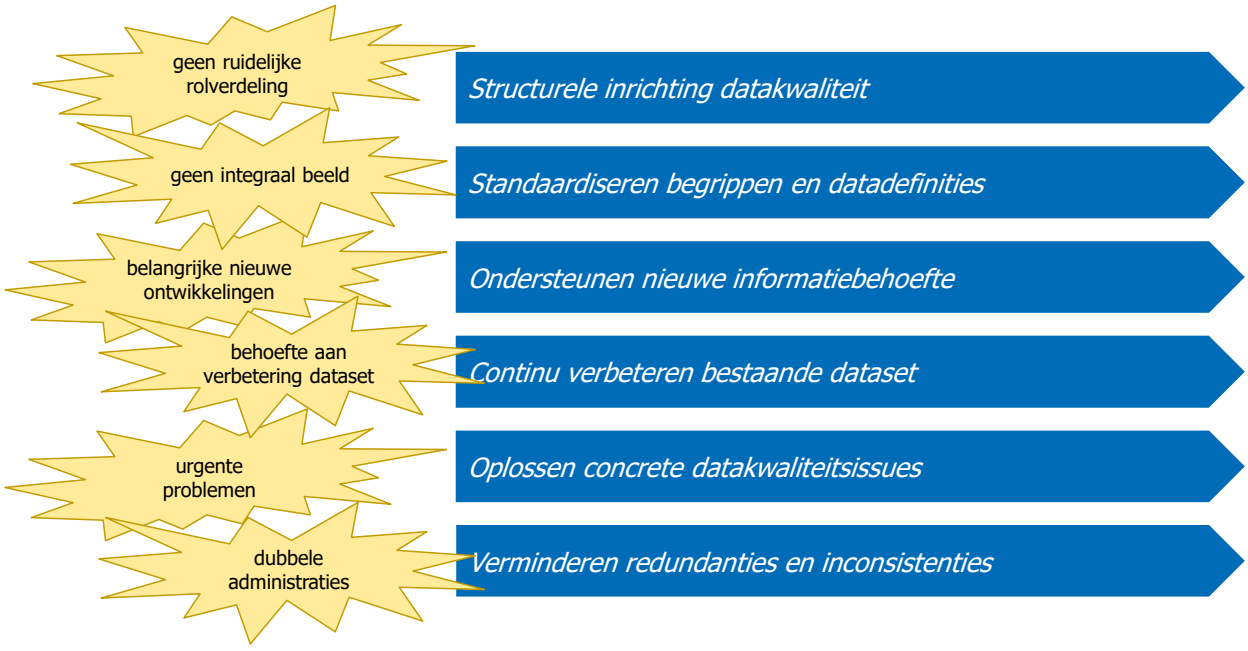
Danny Greefhorst

Er ontstaat steeds meer bewustzijn dat data een belangrijke asset is die expliciete aandacht nodig heeft. Data vraagt om datamanagement. Dat betekent dat er allerlei rollen, processen en systemen moeten worden ingericht rondom data. Het is belangrijk om te beseffen dat dit alles uiteindelijk bedoeld is om de kwaliteit van data aan te laten sluiten bij het gebruik. Datakwaliteit is dus niet slechts een onderdeel van datamanagement; het is het doel van datamanagement. Vanuit dit perspectief kan veel gerichter gewerkt worden aan datamanagement wat daarmee vrijwel synomiem is aan datakwaliteit. Deze whitepaper biedt een praktische beschrijving van wat nodig is om datakwaliteit te realiseren.

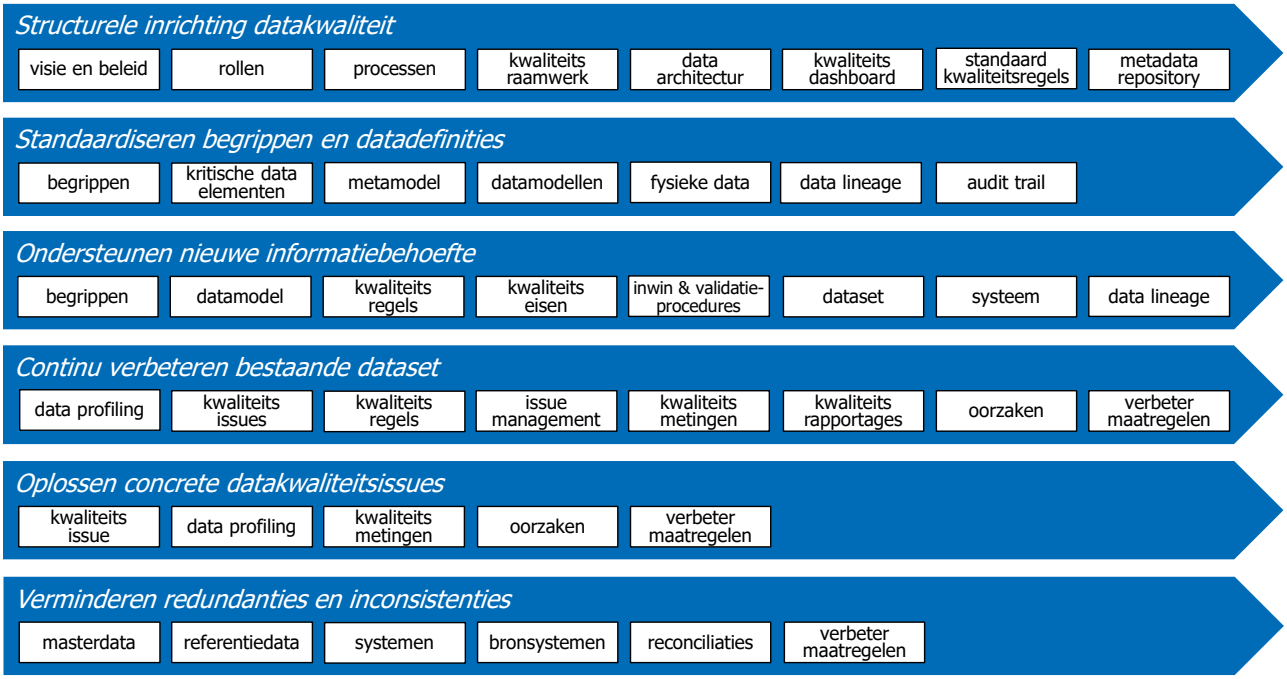
Datakwaliteit

Datakwaliteit is de mate waarin data voldoet aan impliciete en expliciete eisen vanuit het gebruik. Dat gaat met mate over de waarin data nauwkeurig, compleet, consistent, actueel en herleidbaar is. Het gaat bijvoorbeeld over of de getallen in een rapportage wel kloppen. Als dat niet zo is dan is de kans groot dat er verkeerde conclusies worden getrokken of verkeerde besluiten worden genomen. Datakwaliteit is geen absoluut gegeven waarvan je kunt stellen of het wel of niet aanwezig is. Het is aan jezelf om te bepalen wat het betekent in jouw specifieke context. Welke wet- en regelgeving moet aan worden voldaan? Wat is nodig om de doelstellingen te realiseren? Welke eisen stellen de processen aan de data? Wat zijn de regels die niet mogen worden overschreden? Door dit soort aspecten expliciet te maken wordt duidelijk waar de data aan moet voldoen, kan er worden gemeten en gewerkt aan continue verbetering.

Organisaties zijn al op allerlei manieren aan de slag met datakwaliteit, bewust of onbewust. Er zitten in systemen al allerlei controles ingebouwd bij het invoeren van gegevens. Mensen zijn ook al gewend om zelf een oordeel te vellen over data die ze zien. Als ze een fout zien koppelen ze het terug aan de persoon die de data heeft gemaakt. Het is afhankelijk van de organisatiecontext welke verbeteringen mogelijk zijn in het omgaan met datakwaliteit. Er zijn een aantal logische routes te definiëren voor dit soort verbeteringen. Deze whitepaper beschrijft dit soort logische routes. Onderstaand figuur geeft een overzicht. De routes kunnen deels parallel worden ingezet.



Binnen elk van deze routes zijn allerlei stappen en producten relevant. De volgende figuur geeft een globaal overzicht van de termen die daarbij relevant zijn per route. Elk van deze routes en de bijbehorende termen worden in de vervolghoofdstukken in meer detail beschreven.



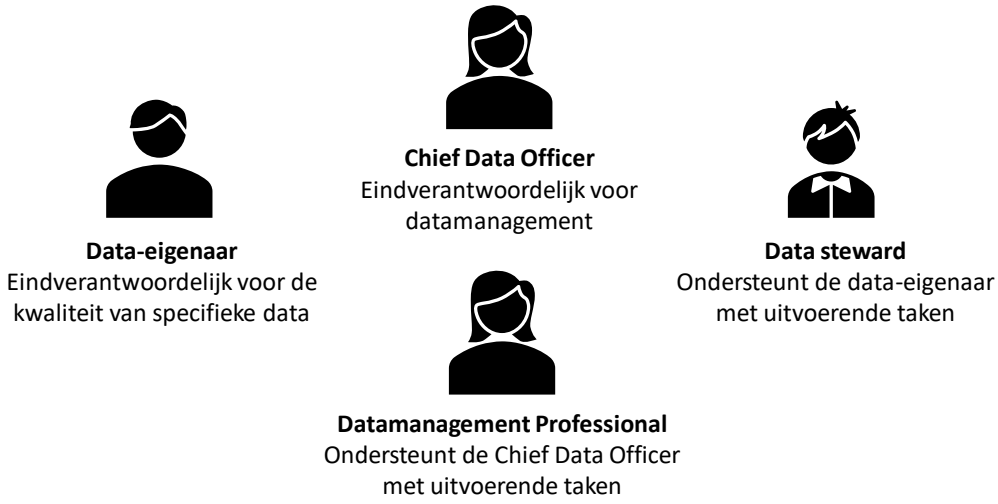
Structurele inrichting datakwaliteit

Als de verantwoordelijkheid voor data niet is belegd dan is de kans groot dat dit een direct effect heeft op de kwaliteit van data. Een structurele inrichting van datakwaliteit start dan ook met het benoemen wie, welke verantwoordelijkheid heeft m.b.t. data. Dat begint met de strategie van de organisatie en vertaalt zich naar een specifieke **visie en beleid** op het gebied van data. De visie geeft aan op welke manier data bijdraagt aan de doelstellingen van de organisatie en wat de belangrijkste eisen zijn die gesteld worden aan de kwaliteit van de data. Wet- en regelgeving kan daarin een bepalende factor zijn. In het beleid worden meer concrete uitgangspunten gedefinieerd die bepalend zijn voor hoe in de organisatie met datakwaliteit moet worden omgegaan.

Het is vervolgens nodig om de **rollen**, taken en verantwoordelijkheden duidelijk te definiëren (data governance). Uitgangspunt daarbij is dat maximaal wordt aangesloten op bestaande rollen in de organisatie. Hierdoor worden datamanagement en datakwaliteit optimaal geborgd in de bestaande organisatie. De basis is de Chief Data Officer (CDO) die eindverantwoordelijk is voor datamanagement in algemene zin. De CDO wordt ondersteund door data management professionals; specialisten op het gebied van datamanagement (inclusief een data-architect). Voor specifieke domeinen zijn data-eigenaren en data stewards nodig die respectievelijk eindverantwoordelijk en uitvoerend zijn voor de data in dat specifieke domein. De data-eigenaar overziet en bewaakt. De data steward gaat concreet aan de slag met specifieke data, verzamelt de eisen, definieert de data, meet de kwaliteit en initieert verbetering. Hoe deze rollen precies samenwerken wordt uitgewerkt in **processen**.

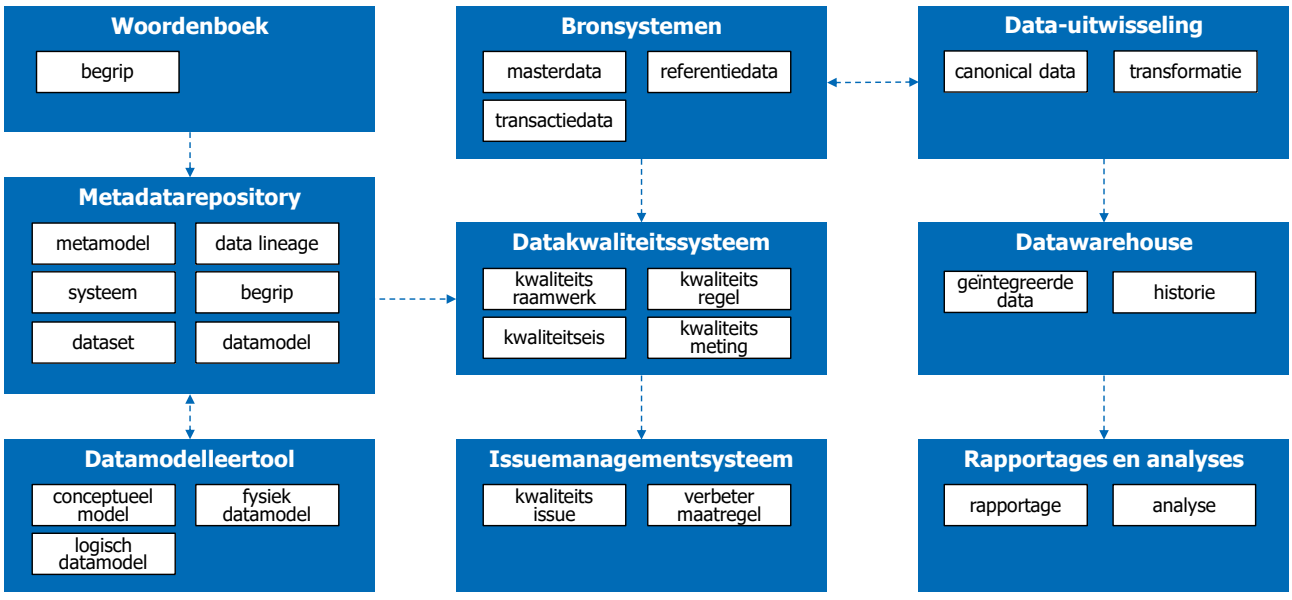
De visie, het beleid, de governancestructuur en de processen zullen moeten worden gecommuniceerd en organisatorisch worden geïmplementeerd. Het vraagt training, opleiding en coaching van medewerkers

zodat zij in staat zijn hun rol uit te oefenen. Het vraagt ook continue aandacht voor het informeren en betrekken van medewerkers.



Om uitdrukking te kunnen geven aan de kwaliteit van data is het nodig om afspraken te maken over de woorden die daarbij worden gebruikt, zoals “compleetheid”, “consistentie”. Dit heet ook wel een **kwaliteitsraamwerk** en de genoemde woorden heten kwaliteitsdimensies. De ISO/IEC 25012 en 25024 standaarden bieden hiervoor een goede basis. In bijlage A is een voorbeeld kwaliteitsraamwerk opgenomen zoals dat is opgesteld in de context van de Omgevingswet. Het biedt een standaard lijst van kwaliteitsdimensies en meer specifieke indicatoren waarbinnen kwaliteit kan worden gedefinieerd en gemeten. Het kan gebruikt worden als checklist voor het identificeren van kwaliteitsissues, kwaliteitseisen, of kwaliteitsregels.

Een **data-architectuur** verdiept visie en beleid naar de gewenste inrichting van systemen op het gebied van data. Het maakt duidelijk welke systemen nodig zijn, welke data en functionaliteiten daarbij horen en hoe ze met elkaar zouden moeten integreren. Dat gaat onder meer over de verschillende soorten databases en hun onderlinge data-uitwisseling. Het gaat ook over systemen die specifiek nodig zijn om datamanagement te ondersteunen. Onderstaand figuur geeft een overzicht van een dergelijke data-architectuur.

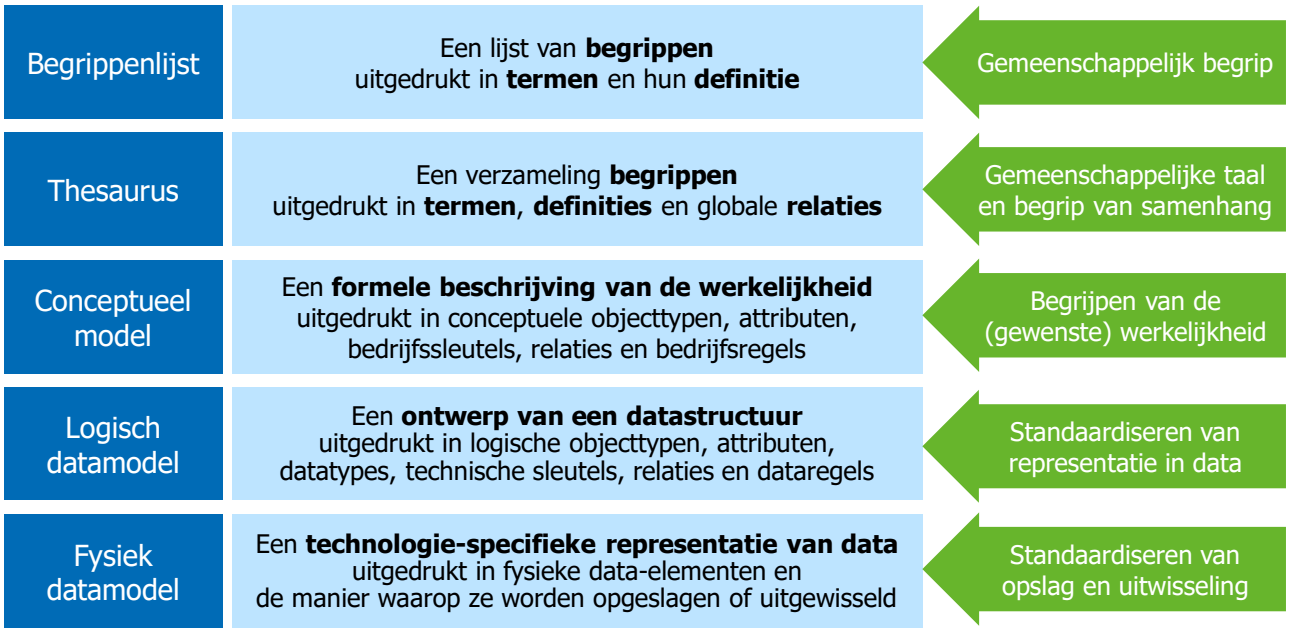


Voor datakwaliteit is een datakwaliteitssysteem belangrijk dat de kwaliteit van data in bronsystemen controleert en inzichtelijk maakt in een **kwaliteitsdashboard** en rapportages. Veelvoorkomende kwaliteitsregels kunnen worden voorgedefinieerd als **standaard kwaliteitsregels**. Geconstateerde problemen met datakwaliteit worden vastgelegd in een issuemanagementsysteem. In dit systeem vindt de verdere afhandeling plaats en wordt gezocht naar verbetermaatregelen.

Een belangrijke basis in de data-architectuur is het vastleggen van metadata in een **metadatarepository.** In deze repository komen definities van systemen, datasets, begrippen, datamodellen en de datastromen (data lineage) bij elkaar. Het wordt gevoed door het woordenboek en een datamodelleertool. De data-architectuur zorgt ervoor dat systemen betekenisvol data kunnen uitwisselen en dat systemen, rapportages en analyses bouwen op goede data.

Standaardiseren begrippen en definities

Het is belangrijk om de werkelijkheid van de organisatie goed te begrijpen voordat je deze in data probeert te verpakken. Dat is met name relevant op het moment dat er onvoldoende organisatiebreed inzicht is in data en wat deze precies betekent. Zonder een dergelijke betekenis kun je niet spreken over de kwaliteit van data. Er moeten dus woorden worden gegeven aan de dingen in de organisatie-werkelijkheid en deze woorden moeten worden gedefinieerd. Dat zijn definities op een taalniveau die nog niets te maken hebben met data. De resulterende **begrippen** moeten worden vastgelegd in een woordenboek. Dat kan gewoonweg een lijst van begrippen zijn of een thesaurus waarin ook de globale onderlinge relaties tussen de begrippen zijn beschreven.



Een standaard lijst van begrippen is een belangrijke basis, maar niet voldoende om de data-uitwisseling tussen systemen te standaardiseren. Juist in de uitwisseling tussen systemen zullen kwaliteitsregels moeten worden gecontroleerd. Een conceptueel model helpt bij het begrijpen van het domein waarin de data wordt uitgewisseld. De data die relevant is om uit te wisselen en de kwaliteitsregels die daarbij gelden worden beschreven in applicatie-onafhankelijke logische **datamodellen** (ook wel: canonical datamodellen). In bijlage B zijn richtlijnen beschreven om de kwaliteit van een datamodel te borgen en beoordelen. Omdat

het opstellen van datamodellen veel inspanning kan kosten is het nodig om te prioriteren. Benoem de **kritische data-elementen**; de elementen die verschil maken en essentieel zijn in formele rapportages.

Er zullen afspraken nodig zijn voor het vastleggen van de datamodellen. Deze afspraken gaan onder meer over de te gebruiken tools, modelleertaal, vast te leggen eigenschappen en datastructuur. Dit vraagt een **metamodel** waarin dit soort afspraken zijn beschreven. Het is het datamodel van de metadatarepository. Binnen de Nederlandse overheid is hiervoor het Metamodel voor Informatiemodellen gedefinieerd, dat uitgaat van datamodellen die zijn opgesteld in UML. Andere organisaties kunnen deze ook gebruiken of ervoor kiezen een meer specifiek metamodel op te stellen, dat rekening houdt met de meer specifieke eisen en wensen van de organisatie. Het kan bijvoorbeeld relevant zijn om meer specifieke eigenschappen voor kwaliteitseigenschappen van de data. Denk bijvoorbeeld aan eigenschappen om eisen m.b.t. de precisie en nauwkeurigheid van attributen te beschrijven.

Er is toenemend behoefte aan informatie over de herkomst van data. Dit is te zien als onderdeel van de kwaliteitsdimensie “herleidbaarheid” en daarmee ook integraal onderdeel van datakwaliteit. In de meest eenvoudige vorm wordt informatie over de herkomst opgenomen in de data zelf. Voor data-elementen waarvoor dat relevant is kan worden opgenomen wie de data heeft gecreëerd, op welk moment en op welke wijze. Als er bijvoorbeeld een specifiek rekenmodel is gebruikt om een bepaald gegeven te berekenen dan kan dat worden opgenomen in de data zelf. Ook andere belangrijke transformaties in tijd kunnen relevant zijn om specifiek op te nemen bij de data zelf. Er kan ontstaat zo een **audit trail** bij gegevens waarin de gehele totstandkoming inzichtelijk wordt.

In specifieke sectoren zoals in de financiële sector worden er nog hogere eisen gesteld aan de herleidbaarheid van gegevens. Er wordt in die context ook wel gesproken over “**data lineage**”. Dat vraagt dat fysieke data op attribuutniveau is gerelateerd aan het bijbehorende logische datamodel en/of hoger liggende begrippen. Het is daarnaast ook nodig om de stroom en transformatie van data op attribuutniveau tussen systemen inzichtelijk te maken. Voor een deel kan dit handmatig worden gedefinieerd, maar voor het in kaart brengen van meer fijnmazige datastromen kan dit eigenlijk alleen op een geautomatiseerde manier. De transformaties om te komen tot het datawarehouse, datamarts en rapporten moeten hiervoor geautomatiseerd worden omgezet tot de relevante metadata.

Ondersteunen nieuwe informatiebehoefte

Nieuwe ontwikkelingen vragen inzicht in de specifieke doelstellingen, eisen en wensen die betrekking hebben op de data. De gebruiksdoelen zijn daarbij leidend. Verschillende gebruiksdoelen stellen andere kwaliteitseisen. Op basis van de gebruiksdoelen kunnen de informatiebehoeften worden geïnventariseerd. Dat vraagt minimaal een globaal begrip van de processen. Zoals eerder beschreven kunnen informatiebehoeften worden uitgedrukt in termen van **begrippen** en verder verfijnd in een conceptueel model. Een verdieping tot een **logisch datamodel** maakt duidelijk welke specifieke data en datastructuur noodzakelijk is. Onderdeel van dit datamodel zijn **kwaliteitsregels**. Deze regels worden gebruikt bij het creëren of ontvangen van de gegevens.

Voor het meetbaar maken van datakwaliteit is het verder belangrijk dat de **kwaliteitseisen** expliciet worden gemaakt. Dit zijn concrete normen waaraan de data moet voldoen. Denk bijvoorbeeld aan een specifieke norm m.b.t. de nauwkeurigheid van de data of het percentage van de data dat mag afwijken van een specifieke verzameling kwaliteitsregels. Dat dient per gebruiksdoel te worden onderzocht, omdat

verschillende gebruiksdoelen ook verschillende kwaliteitseisen kunnen stellen. De basis voor de analyse is de lijst van indicatoren in het kwaliteitsraamwerk. Deze kan worden gebruikt als een checklist voor potentieel relevante aspecten. Waarschijnlijk is maar een deel van de indicatoren relevant voor een specifieke dataset. De kwaliteitseisen kunnen ook betrekking hebben op reeds bestaande datasets.

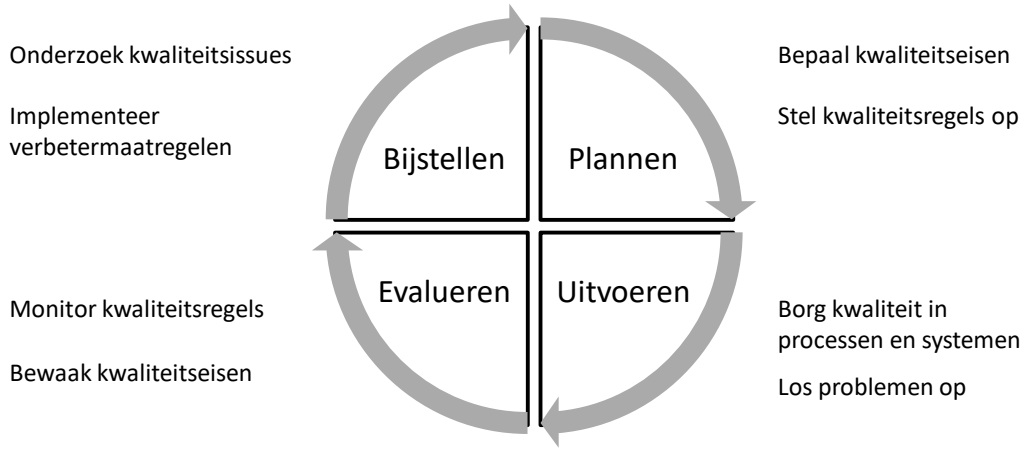
Kwaliteitseisen kunnen ook leiden tot eisen aan de processen waarin de data tot stand komen. In het algemeen kun je dit **inwin- en validatieprocessen** noemen. Een voorbeeld van een dergelijk proces is het proces voor het identificeren van klanten. Dit is op dit moment actueel in het kader van de wet ter voorkoming van witwassen en financieren van terrorisme (WWFT). Het is in dit kader onder meer belangrijk dat de namen van klanten overeenkomen met wat er in hun paspoort staat. Dit betekent dat in het identificatieproces het paspoort een belangrijke rol speelt. Er kunnen extra validaties nodig zijn om voldoende vertrouwen te hebben in de ingewonnen data. In de kader van het voorbeeld kan dat bijvoorbeeld gaan over het controleren van de authenticiteit van het paspoort.

Een nieuwe informatiebehoefte leidt in veel gevallen tot een nieuwe **dataset**. Een dataset is een logische groepering van data-elementen. Een dataset wordt ontsloten via een **systeem**. Dat kan een applicatie zijn of simpelweg een database. Al dit soort dingen moeten kenbaar worden gemaakt in de metadata-repository. In deze repository is bij de datadefinities dus ook bekend in welke datasets zij zijn gebundeld en welke systemen deze datasets aanbieden. Voor de nieuwe dataset is het ook nodig dat de relevante **data lineage** metadata worden vastgelegd. Hierdoor is duidelijk hoe de fysieke data is gerelateerd aan het datamodel en de begrippen en hoe de data stroomt en wordt verwerkt.

Het is belangrijk dat datasets goed vindbaar zijn. De metadata van datasets kan daarvoor ook in andere catalogi moeten worden gepubliceerd. Voor overheidsdata is publicatie op data.overheid.nl en mogelijk ook in het Nationaal Geo Register (www.nationaalgeoregister.nl) van belang. Om afnemers van de data goed te faciliteren is het belangrijk dat de metadata op een toegankelijke manier is beschreven en dat er ook informatie in te vinden is over de kwaliteit van de data. Deze metadata moet zoveel mogelijk inzicht geven in wat een afnemer van deze kwaliteit kan verwachten. Het is bij voorkeur gebaseerd op metingen, alhoewel de metadata typisch minder dynamisch van aard is dan een kwaliteitsdashboard.

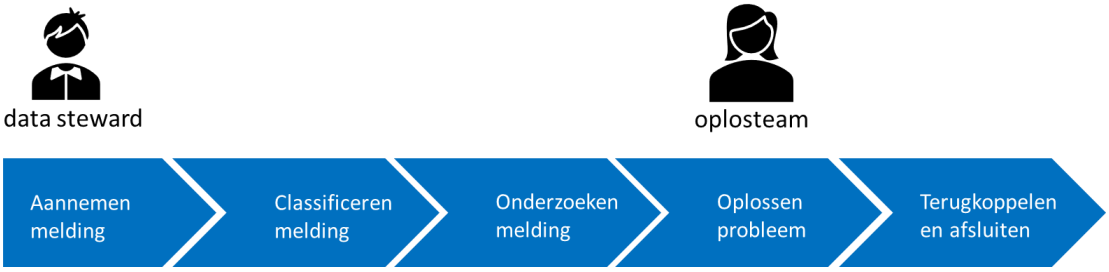
Continu verbeteren bestaande dataset

Datakwaliteit is geen éénmalige inspanning, maar een continu verbeterproces. Het kost nu eenmaal tijd om zaken te verbeteren. Als je stopt na een eenmalige inspanning, dan zul je zien dat de kwaliteit in de loop van de tijd weer terugloopt. Je zou dan ook moeten denken in termen van een kwaliteitscyclus: plannen, uitvoeren, evalueren en bijstellen (zie ook figuur op de volgende pagina). Onderdeel van het plannen is het opstellen van datadefinities, kwaliteitsregels en kwaliteitseisen. Voor bestaande datasets zijn er vaak nog geen kwaliteitsregels en kwaliteitseisen. Het expliciet maken van deze regels en eisen is daarmee een belangrijk aspect van het verbeteren van de kwaliteit van een bestaande dataset.



Voordat wordt gestart met het definiëren van regels en eisen is het gewenst een eerste gevoel te creëren bij de huidige kwaliteit van de dataset. Daarvoor kan gebruik worden gemaakt van **dataprofiling**. Dat is een statistische analyse die inzicht geeft in de huidige data en hun karakteristieken. Het kan voor elke kolom in een tabel de frequentie van waarden bepalen en daarmee inzicht geven in het datatype en gebruik van elke kolom. Het kan ook inzicht geven in minimum, maximum en gemiddelde waarden. Over kolommen heen kan het afhankelijkheden tussen data en overlappende waarden inzichtelijk maken. Op deze manier kan dataprofiling een eerste waardevol inzicht geven in **kwaliteitsissues**. Het identificeert ook mogelijk waardevolle **kwaliteitsregels** (namelijk de regels die de huidige problemen meten) en geeft een beeld bij de **kwaliteitseisen** die gesteld kunnen worden.

Een andere belangrijke bron voor het identificeren van kwaliteitsissues is een **issuemanagement** proces. Dit proces stelt gebruikers van data in staat om problemen die zij ervaren met de data te melden. Hiervoor moet een loket worden ingericht alsook een systeem voor het vastleggen, afhandelen en volgen van issues. In de context van de Nederlandse overheid wordt dit terugmelden genoemd en een dergelijk proces is voor alle basisregistraties ingericht. Het is eigenlijk niet zo belangrijk wie kwaliteitsissues (terugmeldingen) precies indient. Alle meldingen zijn signalen die het onderzoeken waard zijn. Dit onderzoek ligt in eerste instantie bij de data steward, die vervolgens anderen kan inschakelen voor meer diepgaand onderzoek. Melders van kwaliteitsissues verwachten wel dat ze terugkoppeling krijgen over voortgang en oplossing van de issues.



Er zou ook een periodiek proces van **kwaliteitsmeting** moeten zijn om kwaliteitsissues te signaleren. In praktische zin bestaat deze minimaal uit een geautomatiseerd proces dat periodiek controleert of de data voldoet aan de kwaliteitsregels en het resultaat beschikbaat stelt in **kwaliteitsrapportages**. Deze rapportages zouden inzicht moeten geven in alle records waarin afwijkingen van de kwaliteitsregels worden geconstateerd zodat gericht kan worden gewerkt aan verbetering. Ze zouden verder ook inzicht moeten geven in de mate waarin aan de normen zoals beschreven in de kwaliteitseisen wordt voldaan.

Hierdoor kan ook op managementniveau worden gestuurd. Aanvullend op dit standaard proces van kwaliteitsmeting kunnen ook andere metingen worden verricht. Denk bijvoorbeeld aan een meting door een onafhankelijke derde in de vorm van een audit. Dit voorkomt dat de slager zijn eigen vlees keurt.

Op basis van kwaliteitsissues is nader onderzoek nodig om te bepalen wat de **oorzaken** zijn die eraan ten grondslag liggen. Sommige issues zijn eenvoudig van aard en kunnen direct door een data steward worden opgelost. Andere issues zijn groter van aard en de noodzakelijke **verbetermaatregelen** moeten mogelijk worden gerouteerd naar meer specifieke oplosteams. Hele grote issues kunnen zelfs tot projecten leiden, waarvoor ook business cases noodzakelijk kunnen zijn. Kwaliteitsissues moeten in het juiste perspectief worden geplaatst. Dat betekent dat ze moeten worden beschouwd in relatie tot de gestelde kwaliteitseisen, de business impact en de kosten voor het oplossen.

Oplossen concrete datakwaliteitsissues

Er kunnen urgente **issues** zijn met de kwaliteit van een specifieke dataset. Voor een belangrijk deel zijn de activiteiten in een dergelijke situatie vergelijkbaar met die voor het continu verbeteren van een dataset. De urgentie leidt er echter toe dat er sneller geschakeld moet worden en dat er geen tijd is voor het vooraf definiëren van kwaliteitsregels en kwaliteitseisen. Resultaten van eerdere kwaliteitsmetingen zijn wel relevante informatie die een analyse van het probleem kan ondersteunen. Er zal echter zo snel mogelijk moeten worden gewerkt aan een oplossing.

Dataprofiling kan helpen bij urgente kwaliteitsissues omdat dit snel inzetbaar is en snel inzichten geeft. Er kunnen ook meer specifieke **metingen** plaatsvinden om meer zicht te creëren op het probleem. Voor urgente issues geldt nog sterker dan voor andere issues dat als een structurele oplossing niet tijdig kan worden geboden het creëren van een korte termijn oplossing belangrijk is. Dat zorgt ervoor dat de kraan zo snel mogelijk dicht wordt gedraaid en er geen nieuwe issues ontstaan. Er kan vervolgens worden gewerkt aan structurele **verbetermaatregelen**.

Het onderzoeken van de precieze oorzaak van een probleem kan ook veel tijd kosten. Oorzaken kunnen op allerlei vlakken liggen. Veelvoorkomende oorzaken van kwaliteitsissues zijn:

- **Foutieve invoer**: Data wordt niet correct ingevoerd door gebruikers en niet geautomatiseerd gecontroleerd op fouten.
- **Onduidelijke verantwoordelijkheden**: Het is niet duidelijk wie verantwoordelijk is voor het creëren, controleren, bewerken, verbeteren of verstrekken van data.
- **Verschillen en onduidelijkheid in betekenis**: Data die worden gecombineerd zijn ontstaan in verschillende contexten, waarin de betekenis van de data niet expliciet is gemaakt en waardoor ze niet zomaar te combineren zijn.
- **Redundante opslag en beheer**: Data wordt redundant opgeslagen en beheerd in meerdere bronnen, die onderling niet volledig consistent zijn of niet dezelfde actualiteit hebben.
- **Andere gebruiksdoelen**: Data wordt gebruikt voor een gebruiksdoel waarvoor het niet voor ontworpen en ingewonnen is.

Hoe om te gaan met redundantie in opslag en beheer van data wordt uitgewerkt in het volgende hoofdstuk.

Verminderen redundanties en inconsistenties

Gelijksoortige data worden vaak op verschillende plaatsen geadministreerd, waardoor de kans vrij groot is dat deze administraties inconsistent raken. Wijzigingen in één administratie zijn dan niet bekend in de andere administratie of zijn tegelijkertijd ook in de administratie doorgevoerd. Om dit soort problemen te voorkomen of op te lossen is het belangrijk om te werken aan duidelijke bronsystemen. Wijzigingen in data worden alleen doorgevoerd in deze bronsystemen. De data kan om redenen van beschikbaarheid en performance wel worden gerepliceerd, maar wordt niet in andere systemen aangepast.

De belangrijkste soorten data waardoor het belangrijk is om deze slechts op één plaats te administreren zijn **masterdata** en **referentiedata**. Dit zijn beiden relatief stabiele vormen van data die op allerlei plaatsen in het applicatielandschap worden gebruikt. Masterdata is data over objecten met een eigen levenscyclus zoals personen, producten en plaatsen. In administratieve organisaties gaat het vaak over data over klanten. Referentiedata zijn waardelijsten; data die een bepaalde waarde representeren die een bepaald attribuut kan aannemen. Naast masterdata en referentiedata bestaat ook transactiedata. Dat zijn data die ontstaan in de dagelijkse activiteiten van de organisatie, zoals verkooptransacties.

Voor zowel masterdata als referentiedata is het belangrijk om bronsystemen aan te wijzen. Voorafgaand daaraan is het wel noodzakelijk om een duidelijk beeld te creëren van het bestaande systeemlandschap en de data die door deze **systemen** worden gecreëerd, gewijzigd of gebruikt. Een dergelijk landschap wordt typisch vastgelegd in een informatie-architectuur. Deze zou in ieder geval grofmazig inzicht moeten geven in de relatie tussen de data en de systemen. De belangrijkste redundanties worden daarin al zichtbaar. In de beschrijving van de gewenste informatie-architectuur kunnen dan **bronsystemen** worden aangewezen.

Het is ook mogelijk om op een meer gedetailleerd niveau inzicht te geven in redundanties. Dit vraagt een gedetailleerde metadata-administratie op attribuutniveau. Door per attribuut in de logische datamodellen aan te geven in welke systemen deze worden vastgelegd ontstaat een dieper inzicht. Per attribuut kan dan ook een bronsysteem worden aangewezen. Afwijkingen tussen huidige en gewenste situaties leiden tot **verbetermaatregelen** om ervoor te zorgen dat alle data uiteindelijk een eenduidig bronsysteem heeft.

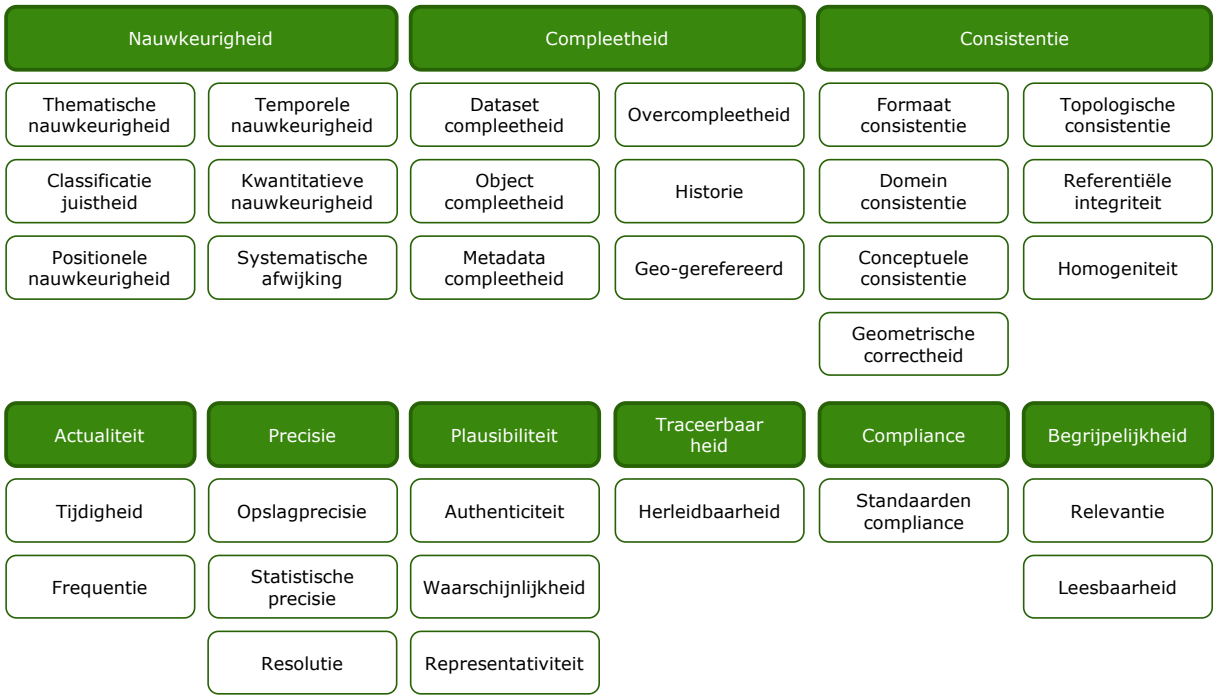
Een belangrijk deel van de verbetermaatregelen is het ervoor zorgen dat systemen relevante data halen uit de bronsystemen door het inrichten van data-uitwisselingen zoals **replicatie**. Deze kunnen worden aangevuld met mechanismen voor **reconciliatie** om te controleren of de data in de systemen nog consistent zijn. Mogelijk moet een deel van de data worden gemigreerd naar het nieuw aangewezen bronsysteem en is het nodig de data te schonen of te transformeren om te voldoen aan het datamodel van het bronsysteem.

Begrippenlijst

- **Attribuut**: Een kenmerk van een objecttype.
- **Begrip**: Een term voorzien van een definitie.
- **Bronsysteem**: Een systeem waarin een bepaald data-element wordt beheerd.
- **Data**: Weergave van een feit, begrip of aanwijzing, geschikt voor overdracht, interpretatie of verwerking door een persoon of apparaat.
- **Datakwaliteit**: De mate waarin een geheel van eigenschappen en kenmerken van data voldoet aan impliciete of expliciete eisen.
- **Data lineage**: informatie over de herkomst en de bewerkingen van data.
- **Datamanagement**: Het ervoor zorgen dat data van voldoende kwaliteit beschikbaar zijn ten behoeve van de bedrijfsprocessen.
- **Datamodel**: Een formele definitie van objecttypen, attributen, relaties en regels.
- **Dataset**: Een verzameling van data die als geheel wordt verwerkt.
- **Datatype**: Het formaat dat wordt gebruikt voor de verzameling van letters, getallen en symbolen om een waarde voor een attribuut weer te geven.
- **Formele historie**: De historie van een object in de registratie.
- **Indicator**: Een meetbaar aspect van een kwaliteitsdimensie.
- **Informatie**: De betekenis van gegevens in een specifieke context.
- **Kwaliteitsdimensie**: Een aspect van kwaliteit waaraan gebruikers van data waarde hechten.
- **Kwaliteitseis**: Een meetbare gewenste eigenschap.
- **Kwaliteitsissue**: Een probleem met de kwaliteit van data.
- **Kwaliteitsmeting**: Een meting die aangeeft in welke mate data voldoen aan kwaliteitseisen en/of daaraan gerelateerde kwaliteitsregels.
- **Kwaliteitsraamwerk**: Een verzameling kwaliteitsdimensies.
- **Kwaliteitsregel**: Een regel die een beperking stelt aan data.
- **Materiële historie**: De historie van de verandering van een object in de werkelijkheid.
- **Meetfunctie**: Een formule die aangeeft hoe een indicator meetbaar kan worden gemaakt.
- **Meetprocedure**: Een beschrijving van de uit te voeren handelingen voor een kwaliteitsmeting.
- **Meetresultaat**: Het resultaat van een kwaliteitsmeting.
- **Metadata**: Data over data.
- **Model**: Een abstractie van de werkelijkheid.
- **Modelelement**: Een onderdeel van een datamodel: een objecttype, attribuut, relatie of regel.
- **Object**: Een fenomeen in de werkelijkheid.
- **Representatie**: De combinatie van waardedomein, datatype en mogelijk een eenheid of karakterset.
- **Objecttype**: Een type van gelijksoortige objecten of data.
- **Systeem**: Een verzameling data en/of functionaliteit.
- **Term**: zie *woord*.
- **Thesaurus**: Een verzameling van begrippen en hun onderlinge relaties.
- **Unified Modeling Language (UML)**: Een standaard voor het definiëren van softwaresystemen.
- **Werkelijkheid**: Een beeld van de echte of hypothetische wereld.
- **Woord**: Het kleinste zelfstandig gebruikte taalelement op het niveau van de spreektaal.

Bijlage A: voorbeeld kwaliteitsraamwerk

In deze bijlage is een voorbeeld kwaliteitsraamwerk beschreven zoals dat is opgesteld in de context van de Omgevingswet. Meer informatie en verdere toelichting is beschreven in het document “Gegevenskwaliteit in de omgevingswet – een verdere handreiking” d.d. 23 februari 2018[1].



Nauwkeurigheid

De mate waarin gegevens de echte waarde in een specifieke gebruikscontext goed weergeven.

Indicator	Beschrijving
Thematische nauwkeurigheid	De mate waarin gegevens die geen betrekking hebben op locatie, tijd of kwantiteit overeenkomen met de werkelijkheid.
Classificatie juistheid	De mate waarin gegevensobjecten zijn geïdentificeerd als het juiste logisch objecttype.
Positionele nauwkeurigheid	De mate waarin locatiegegevens overeenkomen met de werkelijkheid.
Temporele nauwkeurigheid	De mate waarin tijdgegevens overeenkomen met de werkelijkheid.
Kwantitatieve nauwkeurigheid	De mate waarin kwantitatieve gegevens overeenkomen met de werkelijkheid.

[1] zie: https://www.archixl.nl/files/Handreiking_bij_gegevenskwaliteit_in_de_Omgevingswet_1.0.pdf

Systematische afwijking	De mate waarin de resultaten van metingen of berekeningen overeenkomen met de werkelijkheid.

Compleetheid

De mate waarin gegevens gerelateerd aanwezig zijn in een specifieke gebruikscontext.

Indicator	**Beschrijving**
Dataset compleetheid	De mate waarin objecten waarvan het bestaan bekend is aanwezig zijn.
Object compleetheid	De mate waarin attribuutwaarden bij gegevensobjecten aanwezig zijn.
Metadata compleetheid	De mate waarin metadata aanwezig is.
Overcompleet-heid	De mate waarin gegevensobjecten onterecht aanwezig zijn.
Historie	De mate waarin historische gegevens aanwezig zijn.
Geo-gerefereerd	De mate waarin gegevens zijn voorzien van een geo-referentie.

Consistentie

De mate waarin gegevens niet in tegenspraak zijn met andere gegevens in een specifieke gebruikscontext.

Indicator	**Beschrijving**
Formaat-consistentie	De mate waarin gegevens syntactisch correct zijn en daarmee geautomatiseerd te verwerken.
Domein-consistentie	De mate waarin de inhoud en de lengte van attribuutwaarden consistent zijn met hun waardedomein.
Conceptuele consistentie	De mate waarin de combinaties van attribuutwaarden logisch samenhangend zijn.
Geometrische correctheid	De mate waarin gegevens geometrisch correct zijn.
Topologische consistentie	De mate waarin gegevens topologisch consistent zijn.
Referentiële integriteit	De mate waarin gegevens verwijzen naar bestaande gegevens.

Homogeniteit	De mate van variatie in een dataset.

Actualiteit

De mate waarin gegevens recent genoeg zijn in een gebruikscontext.

Indicator	Beschrijving
Tijdigheid	De mate waarin gegevens tijdig zijn geactualiseerd (gemeten of gecontroleerd of ze nog correct zijn).
Frequentie	De mate waarin gegevens met een afgesproken frequentie zijn geregistreerd.

Precisie

De mate waarin gegevens exact of onderscheidend genoeg zijn voor een gebruikscontext.

Indicator	Beschrijving
Opslagprecisie	De mate van detail waarmee gegevens zijn geregistreerd.
Statistische precisie	De mate waarin metingen of berekeningen bij herhaling dezelfde waarde opleveren.
Resolutie	De mate van detail waarmee gegevens worden ingewonnen.

Plausibiliteit

De mate waarin gegevens worden beschouwd als waar en geloofwaardig door gebruikers in een specifieke gebruikscontext.

Indicator	Beschrijving
Authenticiteit	De mate waarin de authenticiteit van gegevens aantoonbaar is.
Waarschijnlijk-heid	De mate waarin gegevens waarschijnlijk zijn voor de situatie.
Representati-viteit	De mate waarin een dataset een goede weergave geeft van het geheel.

Traceerbaarheid

De mate waarin toegang tot gegevens of wijzigingen erin vastgelegd worden in een audit trail in een specifieke gebruikscontext.

Indicator	Beschrijving

Herleidbaarheid	De mate waarin de herkomst, selecties en bewerkingen die hebben plaatsgevonden op de gegevens expliciet zijn vastgelegd.

Compliance

De mate waarin gegevens conformeren aan standaarden, conventies of regelgeving gerelateerd aan gegevenskwaliteit in een specifieke gebruikscontext.

Indicator	**Beschrijving**
Standaarden compliance	De mate waarin gegevens conformeren aan afgesproken standaarden.

Begrijpelijkheid

De mate waarin gegevens eenvoudig gelezen en geïnterpreteerd kunnen worden door gebruikers, en zijn verwoordt in geschikte talen, symbolen en eenheden in een specifieke gebruikscontext.

Naam	**Beschrijving**
Relevantie	De mate waarin de gegevens relevant zijn voor de gebruikscontext of daartoe beperkt kunnen worden.
Leesbaarheid	De mate waarin teksten voor de doelgroep begrijpelijk zijn geformuleerd.

Bijlage B: kwaliteit van datamodellen

Deze bijlage biedt een verzameling richtlijnen voor het borgen en beoordelen van de kwaliteit van logische datamodellen. De informatie in dit hoofdstuk is gebaseerd op algemene literatuur, met name de “data model scorecard” van Steve Hoberman. De richtlijnen zijn ingedeeld in (een deel van) de kwaliteitsdimensies uit de ISO/IEC 25012 standaard.

Nauwkeurigheid

#	Richtlijn
1.1	Het model ondersteunt de gedocumenteerde eisen
1.2	Definities van objecttypen, attributen en relaties zijn een goede weergave van de objecten en hun eigenschappen in de werkelijkheid
1.3	Het model ondersteunt de registratie en/of uitwisseling van de beschikbare gegevens
1.4	Het model abstraheert van informatie die nog onduidelijk is of in de toekomst waarschijnlijk nog zal veranderen

Compleetheid

#	Richtlijn
2.1	Het model beschrijft alle informatie die relevant is in het domein van beschouwing
2.2	Een logisch model bevat alle informatie die nodig is om het conceptuele model te ondersteunen met gegevens, binnen de gekozen afbakening
2.3	Het model blijft binnen de grenzen van het domein van beschouwing
2.4	Het model bevat geen informatie die specifiek is voor het type opslag of uitwisselformaat (geen technologiespecifieke informatie)
2.5	Objecttypen en attributen zijn voorzien van een definitie of een verwijzing naar een begrip in de thesaurus
2.6	Objecttypen in een logisch model zijn voorzien van een identificatie
2.7	Relaties en attributen zijn voorzien van kardinaliteiten
2.8	Voor niet-primitieve typen zijn waardedomeinen gedefinieerd, die kunnen worden hergebruikt voor verschillende attributen
2.9	Een logisch model bevat attributen ter ondersteuning van materiële en formele historie als dat relevant en noodzakelijk is
2.10	Objecttypen die gerelateerd zijn aan een geografische locatie zijn voorzien van een attribuut voor deze locatie

Consistentie

#	Richtlijn
3.1	Attributen representeren slechts één begrip in de thesaurus dat eraan ten grondslag ligt
3.2	Het model bevat geen redundante informatie; het is genormaliseerd tot minimaal de 3e normaalvorm
3.3	Attributen die in meerdere objecttypen worden gebruikt delen dezelfde definitie en overige metadata
3.4	Objecttypen zijn gerelateerd aan andere objecttypen, tenzij duidelijk is dat zij een volledig zelfstandig bestaansrecht hebben
3.5	De definitie en representatie van modelelementen zijn consistent met de beschikbare gegevens
3.6	De definitie van modelelementen verwijst niet naar zichzelf (geen circulaire definitie)
3.7	Modelelementen worden waar mogelijk overgenomen uit andere relevante datamodellen en als zodanig expliciet gemarkeerd
3.8	Vergelijkbare attributen in het model hebben een vergelijkbare representatie
3.9	De belangrijkste ontwerpkeuzes bij het maken van het model zijn expliciet gedocumenteerd, inclusief de daaraan ten grondslag liggende argumenten

Actualiteit

#	Richtlijn
4.1	Het model is bijgewerkt n.a.v. de meest recente versies van brondocumenten die eraan ten grondslag liggen

Precisie

#	Richtlijn
5.1	De definitie van modelelementen is voldoende onderscheidend om te bepalen of instanties van deze modelelementen eraan voldoen
5.2	De naam van modelelementen is specifiek genoeg voor de definitie en om zich te onderscheiden van andere modelelementen die erop lijken
5.3	De identificaties van objecttypen zijn uniek binnen de registratie, stabiel en verplicht
5.4	De identificaties bevatten alleen de minimaal noodzakelijke attributen om gegevensobjecten uniek te identificeren

5.6	Attributen representeren slechts één soort mogelijke waarde (ze hebben maar één betekenis)
5.7	Attributen die direct zijn afleid van andere attributen gaan vergezeld van de afleidingsregel die daarbij moet worden gebruikt
5.8	Attributen leggen informatie niet meer gedetailleerd vast dan wat nodig is om de gedocumenteerde eisen te ondersteunen

Plausibiliteit

#	Richtlijn
6.1	Het model is afgestemd met en geaccordeerd door een representatieve set van gebruikers
6.2	Definities van modelelementen komen overeen met een gangbare of plausibele interpretatie van de naam van het modelelement
6.3	Het model beschrijft informatie die met een hoge mate van waarschijnlijkheid kan worden ingewonnen

Traceerbaarheid

#	Richtlijn
7.1	Modelelementen zijn traceerbaar naar de modelelementen in andere modellen die eraan ten grondslag liggen
7.2	Modellen zijn voorzien van informatie over hoe deze zich verhoudt tot een voorgaande versie van het model (compatibiliteit)

Compliance

#	Richtlijn
8.1	Het model voldoet aan het metamodel

Begrijpelijkheid

#	Richtlijn
9.1	Het model is beschreven in de Nederlandse taal
9.2	De definitie van modelelementen gebruikt alleen termen en afkortingen die eenduidig zijn en begrijpelijk zijn voor de doelgroep of expliciet zijn gedefinieerd

9.3	De definitie van modelelementen is kort en gebruikt geen termen of zinsconstructies die niet nodig zijn om de betekenis van het modelelement te definiëren
9.4	De definitie van modelelementen is voor de lezer duidelijk gescheiden van de toelichting erop
9.5	Het model is opgedeeld in eenvoudig leesbare delen met een duidelijk gedefinieerde onderlinge relatie
9.6	Attributen in een objecttype zijn gesorteerd naar een logische volgordelijkheid en/of gegroepeerd naar logische samenhang
9.7	Modelelementen zijn visueel zo weergegeven en uitgelijnd dat de betekenis niet wordt gehinderd

Over ArchiXL

ArchiXL is een onafhankelijk adviesbureau, gespecialiseerd in enterprise- en informatie-architectuur. Wij adviseren organisaties bij het operationaliseren van hun strategie. De naam ArchiXL is een samenvoeging van "Architectuur" en "XL", waarbij XL staat voor “excelleren”. Wij helpen organisaties om hun doelen te bereiken waardoor zij kunnen excelleren. Onderscheidend daarbij is onze pragmatische en doelgerichte werkwijze. Dat zorgt dat we sterk gericht zijn op het leveren van toegevoegde waarde, passend bij de context van de organisatie. Als specialist op het gebied van architectuur kennen we alle relevante methoden en technieken en weten we als geen ander wat de valkuilen zijn. Onze medewerkers onderscheiden zich door hun communicatieve vaardigheden, resultaatgerichtheid, en hun abstractie- en inlevingsvermogen.

Het is onze passie om de doelmatigheid en effectiviteit van veranderingen en de wijze waarop architectuur en kennis daarbij worden toegepast te verbeteren. Wij denken dat mensen en hun kennis daarin een centrale rol spelen. Het is belangrijk om de specifieke kennis, vaardigheden en talenten van mensen te zien en maximaal in te zetten voor de doelstellingen van de organisatie. De basis daarvoor is een goed gesprek en een goed luistervermogen. In onze visie wordt architectuur nog onvoldoende effectief ingezet om de organisatie te ondersteunen. Symptomen hiervan zijn ontoegankelijke architectuurdocumenten, abstracte modellen die niet aansluiten bij de praktijk en architecten die zich afzonderen van de organisatie. Door kennis te mobiliseren zet je anderen in hun kracht en kom je samen tot grote hoogte.

Onze principes

- Standaard methode – onze aanpak is gebaseerd op standaard methoden en technieken zoals ArchiMate en TOGAF, en daarmee op uitgebreide kennis en ervaring van anderen.
- Hergebruik – organisaties lijken in veel opzichten op elkaar en hergebruik van kennis en architecturen is daarom verstandig.
- Iteratief werken – het is belangrijk om snel antwoord te geven op vragen vanuit de organisatie; dit hoeft niet altijd een volledig antwoord te zijn.
- Concrete en bruikbare resultaten – architectuurproducten moeten direct bruikbaar zijn en waarde opleveren voor de organisatie.
- Samenwerking – veranderen doe je samen, daarmee bundel je ook de kennis en denkvermogen en ontstaat draagvlak voor de verandering.
- ”Just enough” architectuur – architectuurdocumenten moeten bijdragen aan de doelstellingen en niet meer beschrijven dan noodzakelijk.
- Mobiliseren kennis – architecten moeten zich vooral richten op het verzamelen, analyseren, genereren en verspreiden van kennis.

Meer weten?

telefoon: 033-2585545

e-mail: info@archixl.nl

website: http://www.archixl.nl