Limperg Instituut

Op zoek naar onbekende fouten in pensioenadministraties

Anomalie detectie, reversed engineering en gestratificeerde steekproeven in de praktijk

Michael Zuur

Limperg Instituut

Agenda

1. Achtergrond & high level aanpak
2. Analyse op Datakwaliteit
3. Herstelwerkzaamheden

Achtergrond & high level aanpak

Achtergrond

De voorgenomen Wet toekomst pensioenen leidt tot de meest radicale verandering van het Nederlandse pensioenstelsel uit de geschiedenis.

Deze verandering zet opnieuw en duidelijker dan ooit de schijnwerpers op het belang van datakwaliteit. Pensioenfondsen die niet voldoende (aantoonbaar) in controle zijn van hun datakwaliteit, kunnen niet adequaat handelen in de voorliggende transitie en belangrijker: de deelnemers binnen deze fondsen hebben geen duidelijkheid over de juistheid van hun individuele pensioenvermogens na transitie.

FD - 22 dec 2022

De Tweede Kamer stemt in met historische verbouwing van pensioenstelsel. En nu?

PensioenPro - 21 okt 2022

DNB: fondsen moeten kader datakwaliteit goed volgen

DNB is positief over het nieuwe kader datakwaliteit, maar fondsen moeten wel alle stappen goed uitvoeren. Dit stelt de toezichthouder naar aanleiding van de publicatie van de richtlijnen door de Pensioenfederatie.

PensioenPro - 30 sep 2022

Foutieve data bestrijden met algoritmes en handwerk

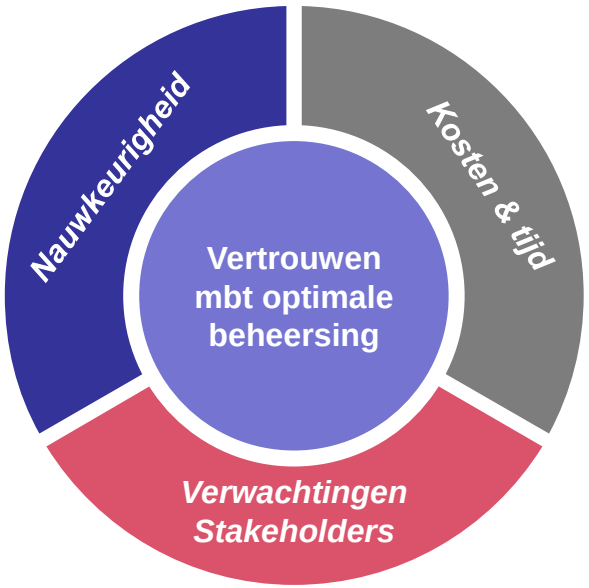
In aanloop naar het nieuwe stelsel zijn pensioenfondsen en uitvoerders druk bezig de datakwaliteit op orde te krijgen. Een geboortedatum van 1900 moet nu al uit de administratie worden gevist.

Francine van Dierendonck APG in 1e kamer: "Het is onmogelijk per deelnemer na te lopen of alles in orde is: de aantallen zijn simpelweg te groot"

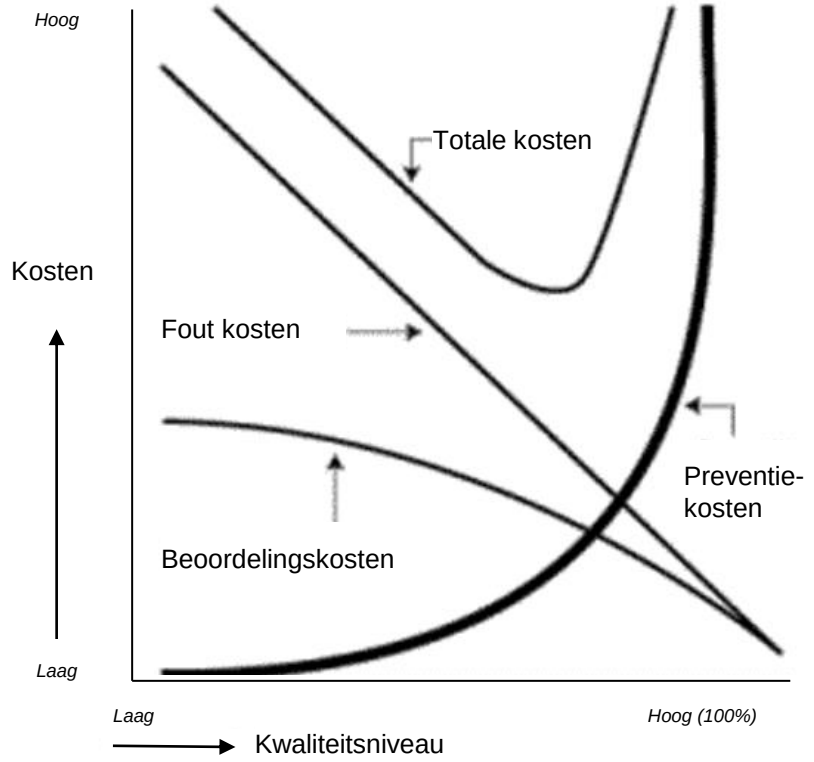
Verschillende mogelijkheden hoe invulling te geven aan het kader

Zekerheid op individueel dossierniveau wenselijk vs mogelijk?

Optimale risicobeheersing is een resultante van het kwaliteitsraamwerk en het vertrouwen van stakeholders waarbij kosten, tijd en beperkingen bepalend zijn voor het maximale kwaliteitsniveau



Op jaarrekening materialiteit gebaseerde controls uit verleden bieden onvoldoende houvast



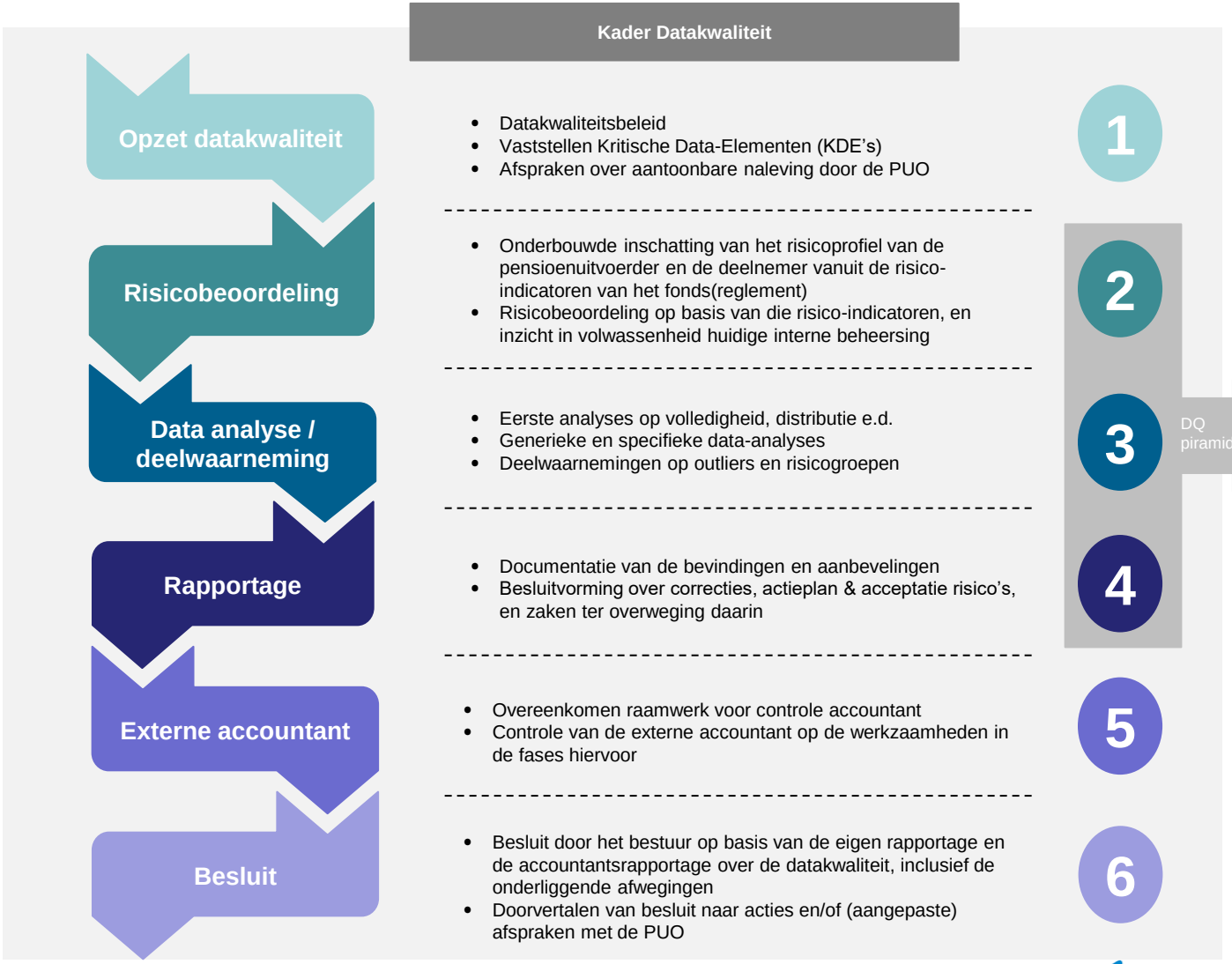
Bepalen vereiste kwaliteit vraagt een evenwichtige afstemming met stakeholders en continue afweging van dilemma's
Transparante proactieve dialoog met stakeholders over vertrouwen en comfort

Het kader Datakwaliteit Pensioenfederatie bestaat uit 6 stappen

In dit licht heeft de Pensioenfederatie een Kader Datakwaliteit opgesteld. Dit kader biedt handvatten aan pensioenfondsbesturen om hun datakwaliteit in beeld te brengen, waar nodig te verbeteren en vast te leggen.

Het Kader is opgebouwd uit 6 stappen, die hiernaast kort staan weergegeven. In de verschillende stappen zijn verschillende technieken en methodieken toe te passen. Hierbij is de uitdaging om deze als PUO zo generiek mogelijk in te richten om te voorkomen dat elk fonds een eigen aanpak vraagt.

Zo blijft of komt u aantoonbaar in controle van uw datakwaliteit en zorgt u ervoor dat uw klanten klaar zijn voor het nieuwe pensioenstelsel!



Analyse op Datakwaliteit

De stappen uit het kader vertaald naar een datakwaliteit piramide

Kader Datakwaliteit

4 **Rapportage**

Het combineren van de verschillende resultaten in een rapportage in vorm van een rapport en eventueel een PowerBI dashboard om totaal inzicht te krijgen in de datakwaliteit van het pensioenfonds.

3 **Deelwaarneming**

Data-analyse

Onderzoek naar juistheid pensioenaanspraken door middel van specifieke analyses (Fase 3.2) zoals forward/reversed engineering (integrale doorrekening), deelwaarnemingen of gestratificeerde steekproeven

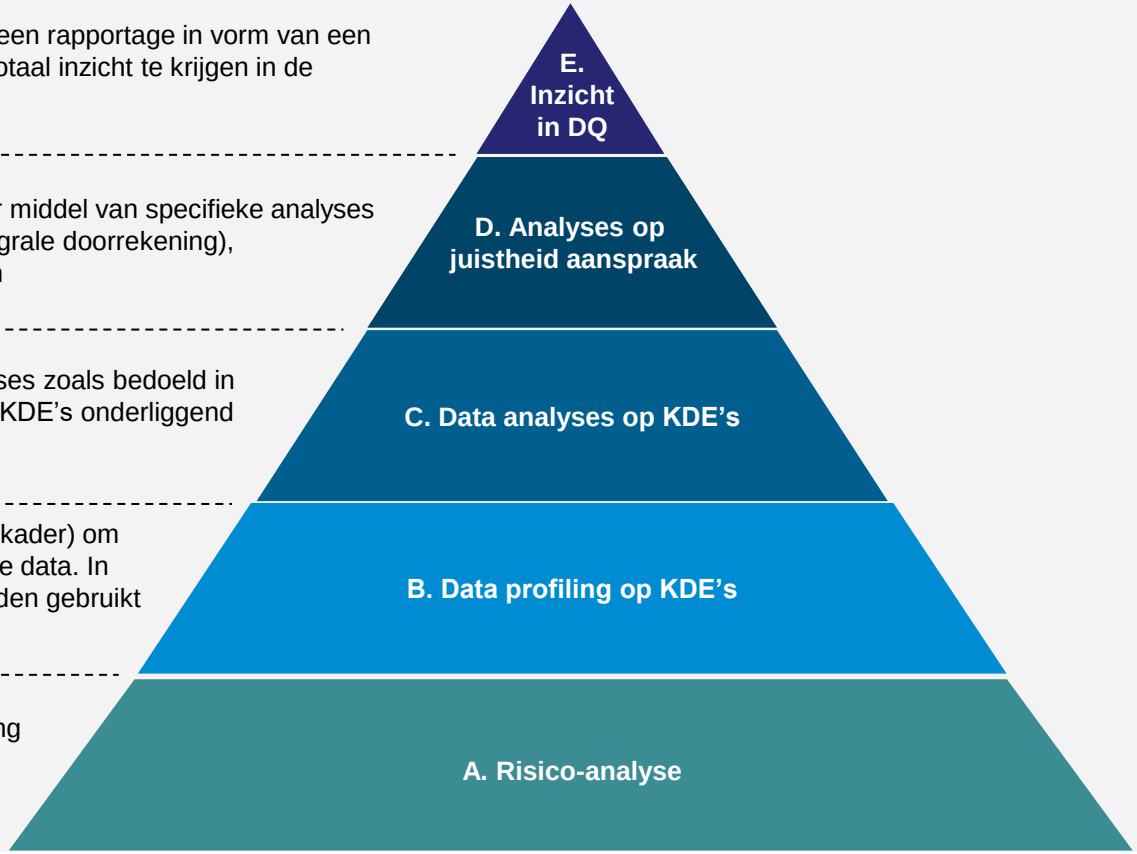
Het uitvoeren van de generieke en specifieke analyses zoals bedoeld in Fase 3.2 van het kader, met daarbij de focus op de KDE's onderliggend aan de aanspraakberekening.

Het uitvoeren van data profiling analyses (Fase 3.1 kader) om inzicht te krijgen in de statistische kenmerken van de data. In deze stap kunnen ook data science technieken worden gebruikt om extra inzicht te verkrijgen.

2 **Risicobeoordeling**

Uitvoeren van de risico-inventarisatie en -beoordeling zoals beschreven in Fase 2 van het Kader.

Datakwaliteit (DQ) piramide



De verschillende lagen van de datakwaliteit piramide bouwen op elkaar voort, om uiteindelijk het gewenste inzicht te kunnen krijgen. Hier kan een onderscheid worden gemaakt tussen de aanspraak KDE en de onderliggende KDE's, en daarmee de manier waarop deze dienen te worden beoordeeld. Daarom zijn hiervoor, in tegenstelling tot in het Kader Datakwaliteit van de Pensioenfederatie, twee verschillende lagen gereserveerd in de datakwaliteit piramide.

Limperg Instituut

Slimme controles door optimale controle mix

Identificatie bekende fouten

Data analyses - Bekende fouten

- Workshops process walkthroughs
- Alignment data management afdeling
- Uitvoeren desktop analyses
- Data analyses bekende fouten
  - Data profiling
  - Consistentie Checks
  - Plausibiliteit Checks

Forward/Reversed engineering

- Zeker stellen merendeel mutaties doordat deze ‘makkelijk’ te verklaren zijn
- Integrale controle high-impact en/of risicovolle elementen

Push left

Push left

Onbekende fouten

Data analyses - Onbekende fouten

- Data profiling
- Trend- / Regressie- / Outlier- / X-check analyses
- Uitsluiten fouten zodat deze niet terugkomen in steekproeven
- Ook voor volledigheid door missende mutaties en X-check met huidige UWV / SVB bestanden elementen

Gestratificeerde steekproeven

- Identificeren onbekende fouten in hoog- / laag- risico segmenten
- Uitspraken over foutkans in overall populatie
- Gebruik inzichten vanuit bekende fouten / al gedane analyses

PowerBI dashboard (incl data preparatie)

Near real-time monitoring, “Geen losse lijstjes”

- Waarden van de fout
- Belasting impact, impact voorziening

Limperg Instituut

Voor het invullen van de verschillende lagen van de piramide zijn verschillende opties op basis van de risico-analyse.

A. Risico-analyse

Scope en type analyse:

Integrale opdeling in risico-segmenten, gebaseerd op kenmerken van het fonds, de deelnemers en de data

Bestaande uit 3 onderdelen:

1. Faciliteren totstandkoming risico-profilering + risicobereidheid per fonds
2. Risico-analyse uitvoering in relatie tot de KDE's in scope
3. Herijken en verrijken uitkomsten risico-analyse door middel van data profiling

B. Data profiling op KDE's

Scope en type analyse:

Integrale data-analyse op KDE's

Basis analyses

Univariate data profiling, te weten:

1. compleetheid,
2. juiste typering,
3. domeinwaarden,
4. (statistische) distributiekenmerken.

Risico gedreven additionele analyses

Op de KDE's met een hoog risico en daaraan verbonden data elementen het uitvoeren van:

- Multivariate verband en outlier analyses
- Anomalie detectie analyses

C. Data analyses op KDE's

Scope en type analyse:

Data analyses op KDE's, sommige toegespitst op de onderliggende KDE's die van belang zijn in de aanspraak berekeningen

55 generieke data analyses uit het kader datakwaliteit van de pensioenfederatie.

Specifieke analyses op basis van risico inventarisatie en/of uitkomsten data profiling.

D. Analyses op juistheid aanspraak

Scope en type analyse:

Combinatie integrale analyse op juistheid van aanspraak berekeningen en gestratificeerde steekproeven per risico segment

Het uitvoeren van deelwaarnemingen op de juistheid van de aanspraak.

1. Specifieke data-analyses op juistheid aanspraken door inzet van forward/backward engineering tool (integrale controle)
2. Waar nodig aangevuld met gestratificeerde steekproeven of deelwaarnemingen

E. Inzicht in Datakwaliteit

Scope en type analyse:

Combinatie van alle onderliggende analyses tot een overall uitspraak op deelnemer niveau

Fonds/PUO klaar voor AUP controle van de accountant.

Rapportage resultaten onderzoek datakwaliteit middels:

1. Rapport, met issues, impact in relatie tot MTA en root causes
2. Aangevuld met inzicht in de datakwaliteit middels een PowerBI dashboard

Data profiling op KDE’s (Stap B)

Data profiling analyses

Voor het uitvoeren van data profiling analyses stelt het Kader Datakwaliteit van de Pensioenfederatie voor dat voor alle KDE’s univariate data profiling analyses worden gedaan. Deze bestaan ondermeer uit:

1. Compleetheid (bijv. geen lege velden),
2. Juiste typering (bijv. datum is een datum),
3. Domeinwaarden (bijv. salaris valt binnen verwachte bereik)
4. (Statistische) distributiekenmerken middels boxplots/histogrammen.

Risico analyse

Indien de risico analyse voor een KDE daar aanleiding toe geeft kunnen deze univariate data analyses verder worden uitgebreid met:

1. Bivariate/multivariate analyses voor het analyseren van relatie tussen variabelen
2. Anomalie detectie, om door middel van machine learning/AI technieken anomalieën in de data te vinden.

Deliverables

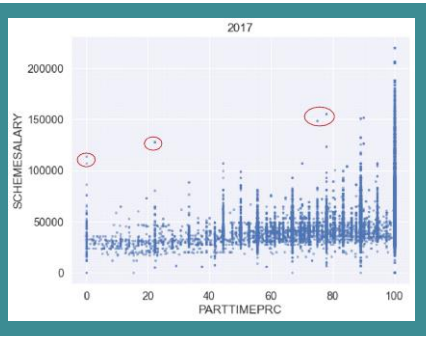
Als deliverable van de data profiling analyses wordt per KDE een rapportage opgeleverd met het inzicht en vervolgstappen vanuit deze analyses.

Univariate analyse

De univariate data profiling analyses geven een eerste inzicht in de distributie van de data. Hiermee kunnen de domeinwaarden van een KDE visueel inzichtelijk worden gemaakt. Daarnaast kunnen statistische outliers worden bepaald middels een inter quartile range

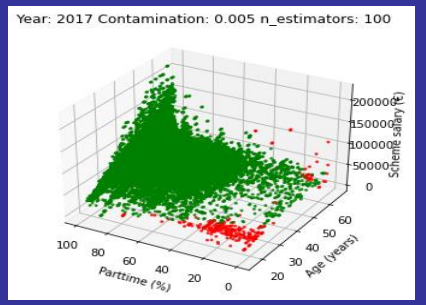
Bivariate & Multivariate

Met bivariate of multivariate data analyses kunnen de verbanden tussen KDE’s inzichtelijk worden gemaakt.



Machine learning Anomalie detectie

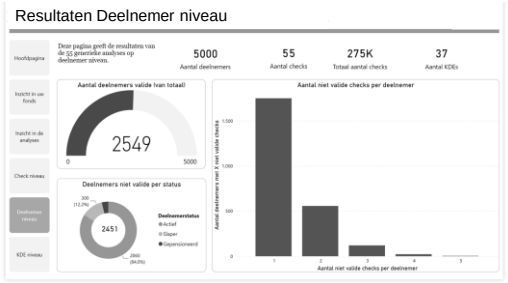
Als sluitstuk van de data profiling analyses kunnen machine learning algoritmen worden gebruikt om onbekende patronen bij een combinatie van meerdere KDE’s te identificeren.



Met behulp van dashboarding kan op een interactieve manier inzicht worden gegeven in de datakwaliteit (Stap E)

Samenvatting op de juistheid van de aanspraak

Een detail dashboard om verder inzicht te krijgen in de juistheid van de aanspraak, door middel van bijvoorbeeld forward engineering.



Het combineren van resultaten

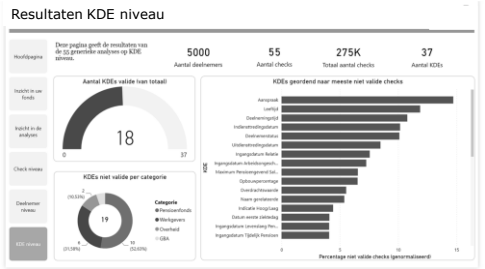
Het eindresultaat van het datakwaliteitsonderzoek is een combinatie van data profiling, data analyses op KDE's en analyses op de juistheid van de aanspraak.

Einduitspraak op deelnemersniveau

"Met xx% betrouwbaarheid hebben meer dan xx% van de deelnemers alleen valide mutaties en geen aandachtspunten in onderliggende KDE's"

Inzicht per KDE

Door in te zoomen op de verschillende KDE's kan ook hier meer detail inzicht in worden gegeven



B. Data profiling op KDE's

"xx% van de KDE's hebben geen aandachtspunten"

C. Data analyses op KDE's

"xx% van de deelnemers hebben geen aandachtspunten in de onderliggende KDE's"

D. Analyses op juistheid aanspraak

"Met xx% betrouwbaarheid hebben meer dan xx% van de deelnemers alleen valide mutaties"

Herstelwerkzaamheden

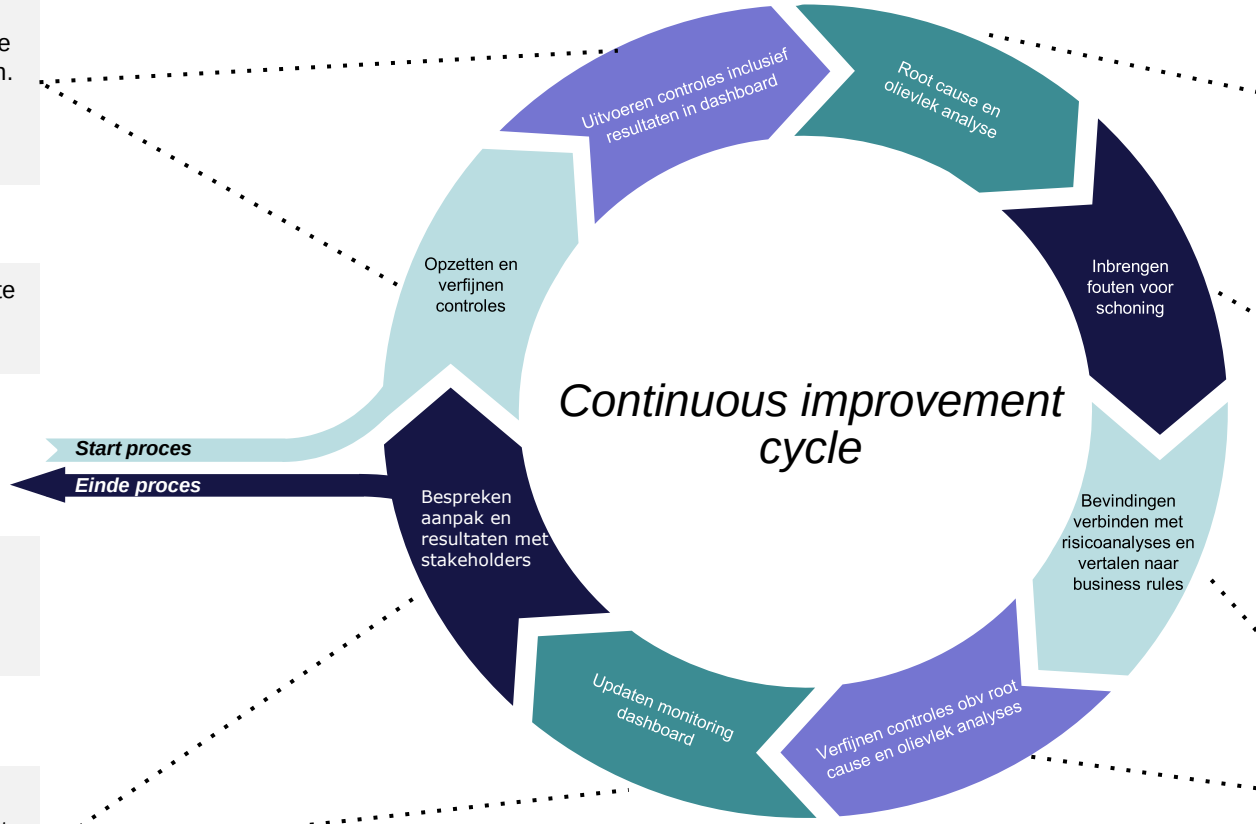
Een analyse op datakwaliteit is niet het eindpunt, er zijn meerdere vervolgstappen nodig om uiteindelijk de datakwaliteit te verbeteren

Het startpunt is een data analyse, zoals een outlier analyse waar initiële bevindingen op KDE's uit voorkomen. Op basis van de eerste inzichten worden de controle-analyses verder verfijnd.

De start van het proces is het vereiste om de datakwaliteit van de KDE's te verbeteren.

Het proces is ten einde wanneer er geen fouten meer worden gevonden en/of wanneer er voldoende betrouwbaarheid is voor het fonds.

Na het inbouwen van de business rules kunnen deze worden geüpdatet in het dashboard. Dit resultaat kan met stakeholders worden besproken om het vervolg te bepalen.



Na het verfijnen van de analyse wordt er op de overgebleven outliers een root cause analyse gedaan om te begrijpen wat er echt aan de hand is. Als uit deze analyse blijkt dat er echt een issue is, dan wordt een olievlek analyse uitgevoerd; is dit data-issue een incident of komt dit (veel) vaker voor?

Op basis van de geïdentificeerde fout wordt data-schoning proces opgestart met data intake en voorstel tot dataschoning. Dit proces volgt het "Auditability by Design'-principe..

Op basis van de gevonden fouten wordt data geschoond. De risico-analyse voor de relevante KDE wordt aangevuld en datakwaliteit business rules aangepast om op de volledige fonds populatie dataset toe te passen. Hierbij kan het push-left principe gehanteerd worden, zodat fouten in het vervolg eerder worden gedetecteerd.

Limperg Instituut

De volgende aanpak voor het herstellen, opschonen en verrijken van data kan dienen als generieke schoningsmethodologie die ook toe te passen is voor latere fondsen

Aanlevering data issue

- Data issues kunnen vanuit geautomatiseerde review of handmatige actie op KDE's geïdentificeerd worden
- Aanlevering data issue op basis van gestandaardiseerd format

Intake data issue

- Controle op correcte invoer en format van het data issue
- Analyse van aangeleverd data issue op:
  - MTA Impact
  - (on)bekende fout
  - relevantie voor invaarproces

Planning schoning data issue

- Root cause analyse van het data issue
- Olievlek analyse
- Schoningsactiviteit op KDE op de backlog-planning met een prioriteit en impactbepaling
- Bepalen handmatige of geautomatiseerde opvolging

Uitvoering schoning data issue

- Voorbereiden schonings-activiteit op data, waar mogelijk via scripts en/of robotisering
- Controle van data schoning met het 'correctie en herstel'-beleid
- Doorvoeren schoning door het schoningsteam van een PUO.

Monitoring schoning data issue

- Monitoring van de voortgang van de schoning van data issues
- Terugkoppeling aan de data owner
- Eventuele communicatie richting deelnemer

Het dataschoning-proces kent een standaard aanpak, waarbij in alle stappen het 'auditable by design' principe wordt toegepast om het proces beheerst en aantoonbaar met documentatie en besluiten vast te leggen