UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Data quality assessment en data cleaning voor procesgerelateerde data

Brecht Steukers

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Niels MARTIN

2021
2022

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Data quality assessment en data cleaning voor procesgerelateerde data

Brecht Steukers

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Niels MARTIN

Het beoordelen van de datakwaliteit van publieke event logs

Brecht Steukers

Faculteit Bedrijfseconomische Wetenschappen, Universiteit Hasselt
Martelarenlaan 42, 3500 Hasselt, België

Process mining is een domein dat de laatste jaren enorm in populariteit is gegroeid. Process mining analyses starten op basis van een event log. Ondanks het grote belang van de datakwaliteit van deze event log, voor de betrouwbaarheid en correctheid van de procesanalyses, wordt vaak niet voldoende aandacht aan de beoordeling ervan besteed. Er worden almaar meer publieke real-life event logs beschikbaar gesteld om proces mining op toe te passen. In deze masterproef wordt de algemene datakwaliteit van deze publieke event logs beoordeeld. Deze beoordeling zal een methode aantonen voor het beoordelen van publieke event logs alsook een overzicht bieden over de algemene datakwaliteit van deze event logs. De beoordeelde event logs worden met elkaar vergeleken in een benchmarking studie. De benchamarking studie zal uitgevoerd worden op vijf diverse publieke real-life event logs van de jaarlijkse *BPI challenge*.

Keywords - *Process mining, Event log, Datakwaliteit, Kwaliteitsbeoordeling*

1 Introductie

In bedrijven worden activiteiten opgedeeld in verschillende bedrijfsprocessen. Data van uitgevoerde bedrijfsprocessen kunnen worden vastgelegd in datasets, die omgevormd kunnen worden tot event logs [23]. Deze event logs bevatten dus procesuitvoeringsdata of event data van de uitgevoerde events [21]. De toenemende hoeveelheid event data zorgt voor veranderingen in *business process management* (BPM) [29]. BPM past een combinatie van kennis in de informatietechnologie met kennis in management wetenschappen toe op operationele bedrijfsprocessen [30]. Het gebruik van de beschikbare event data zorgt voor waardevollere modellering, modelgebaseerde analyses en modelgebaseerde implemenaties van processen. Deze technieken hebben mogelijks het potentieel om te leiden tot eventuele procesverbeteringen [19, 29]. Process mining is een relatief nieuwe onderzoeksdiscipline, die de brug vormt tussen data mining en data analytics aan de ene kant, en procesmodelleren en analyses aan de andere kant [30]. In het BPM onderzoek is process mining uitgegroeid tot een van de belangrijkste thema's en ook de industrie toont veel interesse in dit onderzoeksdomein [30]. Enkele voorbeelden van mogelijke toepassingen van process mining zijn: (1) Het automatisch en efficiënter ontdekken van procesmodellen zonder menselijke modellering, (2) het vinden van bottlenecks (of knelpunten) in bedrijfsprocessen en begrijpen waarom deze bestaan, en (3) het ontdekken en begrijpen van afwijkingen van het procesmodel [19].

De afwezigheid van voldoende kwaliteitsvolle event data is in hedendaags process mining de limiterende factor [6, 17, 19, 27, 30]. De grote hoeveelheid beschikbare data zijn vaak niet juist van structuur om eenvoudig gebruikt te worden voor analyses. Hiervoor zal de data eerst opgeschoond, gefilterd en omgevormd moeten worden tot event logs [19]. De typische structuur van process mining data (event logs) zal bovenop de algemene datakwaliteitsproblemen, bijvoorbeeld ontbrekende data, op zichzelf zorgen voor specifieke event log problemen [10]. Een voorbeeld van een datakwaliteitsprobleem, specifiek voor event logs, is een fout gelogd tijdstip van een event. Hierdoor is het bijvoorbeeld mogelijk dat bij de analyses geacht wordt dat dit event eerder heeft plaatsgevonden dan een ander event, terwijl dit in de realiteit niet het geval was. Net zoals met andere vormen van data analyses, zal de kwaliteit en betrouwbaarheid van de resultaten sterk afhangen van de kwaliteit van de input data. Hier wordt naar gerefereerd als het *garbage in - garbage out* principe [1].

Er worden steeds meer real-life event logs publiek beschikbaar gesteld. Aangezien er bij deze event logs geen toegang is tot extra inputs, bijvoorbeeld domeinkennis, is het belangrijk om vanuit de event logs zelf een beeld te kunnen vormen van de datakwaliteit. In deze masterproef is een empirische studie uitgevoerd om een algemeen beeld te vormen over de datakwaliteit van publieke event logs. Een set van diverse event logs is beoordeeld en vergeleken op basis van de datakwaliteit. Deze beoordeling toont een methode aan om de datakwaliteit van publieke event logs te kwantificeren. Bovendien biedt het ook een overzicht van de datakwaliteit van publieke event logs.

Deze masterproef is als volgt gestructureerd. Sectie 2 bespreekt de onderzoeksvragen en de methodologie van deze paper en verduidelijkt daarenboven de basis concepten van process mining in meer detail. De literatuurstudie in sectie 3 biedt een duidelijk overzicht over de bestaande taxonomieën, datakwaliteitsdimensies en imperfectie patronen om de datakwaliteit van event logs te beoordelen of datakwaliteitsproblemen in deze event logs te identificeren. Sectie 4 beschrijft de uitvoering van de benchmarking studie waarin de event logs met elkaar vergeleken worden op vlak van datakwaliteit. De resultaten en conclusie worden respectievelijk besproken in sectie 5 en sectie 6.

2 Onderzoeksvraag en methodologie

In deze sectie wordt de onderzoeksvraag besproken die beantwoord wordt in deze masterproef (sectie 2.1). Verder zal ook de toegepaste methodologie van zowel de literatuurstudie (sectie 2.2.1) als de empirische studie (sectie 2.2.2) in detail verduidelijkt worden.

2.1 Onderzoeksvraag

De datakwaliteit van event logs heeft een grote invloed op de resultaten van de hierop uitgevoerde proces analyses. Deze masterproef tracht een methode aan te reiken om publieke event logs te beoordelen op vlak van datakwaliteit en een algemeen overzicht te bieden over de kwaliteit van deze event logs. Meer specifiek wordt de volgende onderzoeksvraag beantwoord: *'Wat is de kwaliteit van publieke real-live event logs?'*.

Om deze onderzoeksvraag te beantwoorden, werd eerst onderzocht welke taxonomieën, datakwaliteitsdimensies, en imperfectie patronen ontwikkeld zijn in de literatuur om de datakwaliteit van event logs te beoordelen of om datakwaliteitsproblemen van event data te identificeren. Met behulp van deze literatuurstudie is bekeken hoe de datakwaliteit van event logs gekwantificeerd kan worden. Bij publieke event logs, is er vaak geen andere input beschikbaar dan de event log zelf. Daarom is het beoordelen van de event log gebeurd op basis van enkel de event log, dus zonder gebruik van externe inputs zoals een domeinexpert. Verder moeten de toegepaste metrieken in voldoende mate geoperationaliseerd zijn om deze zelf te kunnen toepassen op de event logs.

Op basis van een diverse set van vijf publieke real-live event logs van de *BPI challenge* is de datakwaliteit van publieke event logs beoordeeld en met elkaar vergeleken. De literatuurstudie dient als input om deze event logs te beoordelen en met elkaar te vergelijken in de benchmark studie. Uit de gevonden taxonomieën, dimensies, en metrieken is een set geselecteerd worden die zelf is toegepast op de geselecteerde event logs.

2.2 Methodologie

2.2.1 Literatuurstudie

De literatuurstudie focust op het onderzoeken van, zowel de reeds bestaande dimensies om datakwaliteit te beoordelen, als de bestaande taxonomieën en imperfectiepatronen om datakwaliteitsproblemen te identificeren. Deze literatuurstudie is voornamelijk gebaseerd op papers uit wetenschappelijke tijdschriften en conferentie proceedings. Deze wetenschappelijke artikelen zijn verkregen via het zoeken in de volgende online databanken: *Uhasselt Discovery*, *Google Scholar* en *Ebscohost*. Deze online databanken bevatten namelijk een grote hoeveelheid aan bronnen die kunnen bijdragen aan het oplossen van de onderzoeksvragen. De literatuurstudie is gestart vanuit hoofdstuk 5, *Data Quality in Process Mining* uit het boek *Interactive Process Mining in Healthcare* [10]. Op basis van de informatie in dit hoofdstuk zijn de initiële zoektermen van deze masterproef bepaald, dewelke verder zijn aangevuld wanneer het onderzoek vorderde. De finale lijst met zoektermen is opgenomen in Tabel 1.

data quality **AND** (event*log **OR** process mining **OR** process-oriented data)
data quality assessment **AND** (event*log **OR** process mining **OR** process-oriented data)
(imperfection **OR** data*issue **OR** taxonomy) **AND** (event*log **OR** process mining **OR** process-oriented data)
data quality dimensions **AND** (event*log **OR** process mining **OR** process-oriented data)

Tabel 1: Initiële zoekwoorden lijst

Bij het opzoeken van artikelen om op te nemen in de masterproef in hiervoor vernoemde online databanken, werden telkens al de volgende criteria toegepast door deze in te stellen bij de zoekfilters van elke zoekopdracht. Enkel het laatste item in de lijst werd individueel per bron bekeken om te kijken of de bron relevant is voor deze masterproef:

- Alle bronnen moeten *peer reviewed* zijn.
- Enkel bronnen geschreven in het Engels of Nederlands worden bekeken.
- Enkel *journal papers*, *conference papers*, masterproeven, en boeken worden bekeken.
- De bron bevat informatie over het opschonen en beoordelen van datakwaliteit of over het identificeren van datakwaliteitsproblemen. Daarenboven worden ook bronnen opgenomen die bijdragen aan het uitleggen van de geziene concepten in deze masterproef. Zo kan het zijn dat bepaalde concepten in de literatuurstudie op zichzelf niet verduidelijkend genoeg zijn om te begrijpen.

Bij het opzoeken werden de resultaten gesorteerd op relevantie en werd er gestopt met het analyseren van de bronnen wanneer er 30 artikels op een rij geen relevant artikel gevonden werd. De relevantie werd bepaalt op basis van een eerste screening van de titel en het abstract van de bron. Indien de titel en het abstract niet genoeg informatie bevatten om de relevantie te bepalen, werd de tekst nauwkeuriger bestu-

deerd op basis van de inleiding en de conclusie. Indien op basis van de titel, het abstract, de inleiding en de conclusie de relevantie van de bron nog steeds onduidelijk is, werd een meer gedetailleerde screening toegepast. Een meer gedetailleerde screening bestaat uit het screenen van de tekst door het bekijken van de tussentitels. Indien een bepaalde tussentitel nuttige informatie leek te bevatten, werd de bijhorende tekst gelezen om zo met zekerheid te kunnen bepalen of het artikel relevant is voor de literatuurstudie.

Behalve het opzoeken van bronnen via de hierboven beschreven zoektermen, werden ook artikels bekeken via de referenties. Meer specifiek werd op alle relevante bronnen *backward reference searching* toegepast. De gebruikte referenties werden in de relevante bronnen bekeken om te bepalen of deze extra informatie konden bijdragen aan de literatuurstudie (*backward reference searching)*) [26]. De relevantie van deze bronnen werd achterhaalt op dezelfde methode die toegepast is op de gevonden bronnen in de online databanken.

In de gevonden, geselecteerde bronnen, werden de relevante secties volledig gelezen. Of een sectie relevant is, werd bepaald aan de hand van de tussentitels, net zoals eerder werd beschreven bij de meer gedetailleerde screening. Tijdens het lezen, zijn markeringen en korte notities aangebracht. Bovendien werd een korte samenvatting en structuurschema van elk artikel gemaakt om een duidelijk overzicht te hebben over de gebruikte bronnen.

2.2.2 Empirische studie

De empirische studie bestaat uit een *benchmarking studie* over de datakwaliteit van publieke real-life event logs. De vergelijkende studie is opgezet om de datakwaliteit van verschillende event logs te beoordelen en met elkaar te vergelijken. Deze beoordeling zal een methode aanreiken om event logs te beoordelen alsook een overzicht bieden over de kwaliteit van publieke real-life event logs. De beoordeling van de event logs is uitgevoerd op basis van datakwaliteitsdimensies en metrieken om de datakwaliteit van event logs te beoordelen. Deze zijn geselecteerd op basis van de literatuurstudie van sectie 3. De publieke *real-life* event logs van de *BPI challenge* zijn gebruikt als input voor de benchmarking studie. De stappen om een kwaliteitsvolle benchmarking studie uit te voeren, zijn gebaseerd op de gebruikte methode door Augusto et al. [3]. In dit artikel vergelijken de auteurs aan de hand van de volgende stappen de kwaliteit van meerdere automatische algoritmen die procesmodellen ontdekken [3].

- Methode selectie
- Setup en datasets
- Evaluatie metrieken
- Benchmarking resultaten

De gevolgde stappen van Augusto et al. [3] zijn aangepast aan de context van deze masterproef en hebben geleid tot de volgende vier stappen:

- **Gebruikte tools**: In deze stap worden de gebruikte tools besproken die de datakwaliteit van de event logs zullen beoordelen.
- **Event logs**: Vervolgens wordt de selectie van publieke real-life event logs besproken. In deze stap wordt een introductie gegeven om de geselecteerde event logs te introduceren en worden de verschillen tussen de event logs besproken om de diversiteit van de selectie aan te tonen.
- **Datakwaliteitsdimensies en metrieken**: De datakwaliteit van de gebruikte event logs wordt bepaald aan de hand van een set van datakwaliteitsdimensies die beoordeeld worden met behulp van een set metrieken per dimensie. In deze stap worden de dimensies en metrieken besproken en verduidelijkt.
- **Benchmarking resultaten**: De wordt studie uitgevoerd en worden de resultaten besproken in deze stap. De datakwaliteit per event log wordt besproken en de verschillen tussen de event logs worden vergeleken.

2.3 Terminologie

In event logs wordt historische informatie over de uitvoering van bedrijfsprocessen vastgelegd [23]. Event logs bevatten procesuitvoeringsdata, vastgelegd door informatiesystemen die een bepaald bedrijfsproces ondersteunen [21]. Een event log is een verzameling van events, gebruikt als input voor process mining [30]. Elk event in een event log, verwijst naar een activiteit (een goed gedefinieerde stap in een proces) en is gerelateerd aan een bepaalde case (de entiteit die wordt afgehandeld door het proces dat wordt geanalyseerd, bijvoorbeeld een klanten order) [29]. Een event is een actie die is vastgelegd in de event log, bijvoorbeeld het starten, voltooien of annuleren van een activiteit voor een bepaalde case [30]. De events in een event log worden geordend aan de hand van tijdstippen die toegevoegd worden aan elk event [21]. Een event log kan bovendien ook worden gezien als een verzameling van traces. Elke trace beschrijft de levenscyclus van een bepaalde case in termen van de uitgevoerde activiteiten [29]. Een trace is de opeenvolging van alle activiteiten die gevolgd worden door een case. Een trace kan meerdere cases vertegenwoordigen, die eenzelfde uitvoering van activiteiten hebben ondergaan in het bedrijfsproces. Event logs slaan bovendien vaak aanvullende informatie op over de opgenomen events. Voorbeelden hiervan zijn de resource en de status [29]:

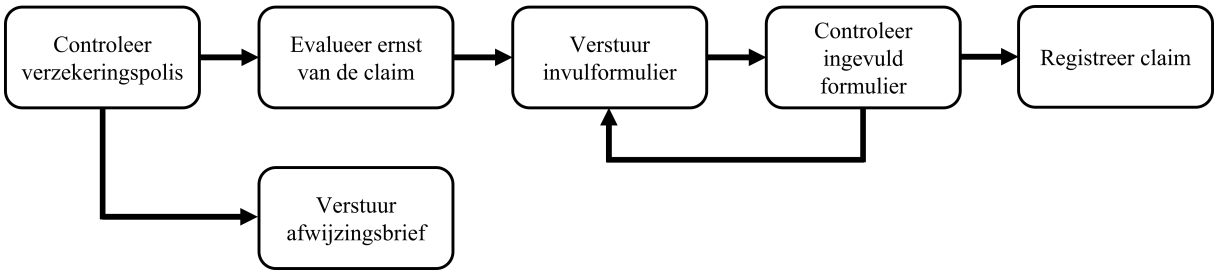
- **Resource**: de persoon, de machine of het softwarecomponent die het event uitvoert
- **Status**: het transactietype van het event (bijvoorbeeld start, voltooiing, pauzeren, hervatten)

Een fictief voorbeeld van een event log is opgenomen in Tabel 2.

Case ID	Activiteit	Status	Resource	Tijdstip
246	Controleer verzekeringspolis	Compleet	367-893	30/05/2022 14:57:36
246	Evalueer ernst van de claim	Compleet	349-873	30/05/2022 15:02:51
247	Controleer verzekeringspolis	Compleet	367-893	30/05/2022 15:02:53
246	Verstuur invulformulier	Compleet	349-873	30/05/2022 15:07:12
248	Controleer verzekeringspolis	Compleet	367-893	30/05/2022 15:10:47
247	Verstuur afwijzingsbrief	Compleet	367-893	30/05/2022 15:11:49
246	Controleer ingevuld formulier	Start	349-873	31/05/2022 09:13:49

Tabel 2: Voorbeeld fictieve event log

Een uitgebreide set van algoritmen is ontwikkeld om (semi-)automatisch inzichten te halen over processen uit event data. Enkele voorbeelden van deze inzichten zijn de ontdekking van de volgorde van activiteiten in een bedrijfsproces aan de hand van de geregistreerde events en de naleving van het proces aan een normatief model. Ook wordt process mining gebruikt bij andere technieken bijvoorbeeld bij het maken van simulaties of bij procesautomatisering [21]. Alle process mining technieken veronderstellen dat event logs zijn opgebouwd zodat elk event verwijst naar een activiteit van het procesmodel en gerelateerd is aan slechts één specifieke case [30]. Om het process mining concept met een voorbeeld verder te verduidelijken, zie het onderstaande (fictieve en versimpelde) procesmodel van een claim verwerkingsproces in Figuur 1.



Figuur 1: Visuele voorstelling van een versimpeld verzekeringsclaim registratieproces

De activiteiten in dit procesmodel zijn: controleer verzekeringspolis, verstuur afwijzingsbrief, evalueer ernst van de claim, verstuur invulformulier, controleer ingevuld formulier, en registreer claim. Een case omvat de verwerking van één specifieke claim, bijvoorbeeld claim 246, die het proces doorloopt. Bij deze case wordt eerst de verzekeringspolis gecontroleerd, dan wordt de ernst van de claim geëvalueerd, een invulformulier verstuurd, het invulformulier gecontroleerd en ten slotte wordt de claim geregistreerd. Bij een andere case, claim 247 bijvoorbeeld wordt de verzekeringspolis gecontroleerd en erna wordt een afwijzingsbrief verstuurd. Een mogelijk event is het controleren van de verzekeringspolis van claim 246. Van dit event zal bovendien extra informatie opgeslagen worden, zoals het tijdstip dat gestart wordt met het controleren van de verzekeringspolis alsook het eindtijdstip. De event log van dit proces zal dus alle events bevatten waarin, voor één specifieke claim, één van de activiteiten van bovenstaand procesmodel uitgevoerd wordt in de periode.

3 Literatuurstudie

Datakwaliteitsproblemen kunnen leiden tot verkeerde resultaten en misleidende statistieken. Om deze problemen te voorkomen, is het essentieel gebruik te maken van kwaliteitsvolle data en op basis hiervan een nauwkeurig inzicht te krijgen in het proces en de juiste beslissingen te nemen [25]. Het beoordelen van de datakwaliteit, alvorens ermee te werken, is hiervoor vereist [12]. Taxonomieën stellen datakwaliteitsproblemen op een gestructureerde manier voor. In deze literatuurstudie wordt een selectie van deze taxonomieën besproken. In sectie 3.3 worden eerst enkele meer algemene taxonomieën besproken, dit zijn classificaties zonder een specifieke focus op process mining. Sectie 3.4 zal taxonomieën van datakwaliteit in event logs bespreken [10]. Nadat de datakwaliteit beoordeeld en de datakwaliteitsproblemen in beeld zijn gebracht, kan dit inzicht gebruikt worden op de datakwaliteit van de event log te verbeteren [10].

3.1 Datakwaliteitsbeoordeling

Event log data van slechte kwaliteit leiden vaak tot verkeerd ontdekte procesmodellen. De kwaliteit van een event log heeft niet alleen invloed op de process mining resultaten [15, 16], maar ook op de keuze van het process mining algoritme [32] (*alpha algoritme* kan je bijvoorbeeld best niet toepassen op real-life event logs [15]). Het is noodzakelijk om de kwaliteit van event logs te beoordelen, voordat er kan worden verder gegaan met het uitvoeren van process mining [28]. De bestaande process mining methodologieën houden desondanks niet specifiek rekening met de kwaliteit van de event logs op het begin van een process mining project. Bovendien zijn er weinig richtlijnen om de datakwaliteit van event logs te beoordelen [2]. Het effect van de datakwaliteit van event logs op de betrouwbaarheid van de resultaten benadrukt de noodzaak van een grondige beoordeling. Op basis van de beoordeling kunnen initiatieven genomen worden om de datakwaliteitsproblemen te verminderen (bijvoorbeeld door aanvullende data te verzamelen) of, indien het niet mogelijk is om een probleem aan te pakken, kan kennis over de aanwezigheid ervan worden meegenomen bij het uitvoeren van proces mining analyses [10].

Het beoordelen van de datakwaliteit van event logs kan worden ondersteund door de hiervoor voorziene tools. Tools identificeren de aanwezige datakwaliteitsproblemen zodat deze in rekening gebracht kunnen worden bij het uitvoeren van de analyes [10]. Eén van deze tools is DaQAPO. DaQAPO (*Data Quality Assessment for Process-Oriented data*) is een open source R-pakket om systematisch de kwaliteit van event logs te beoordelen [10]. R is een programmeertaal die het manipuleren van data en het uitvoeren van statische analyses mogelijk maakt aan de hand van een uitgebreide set functionaliteiten [13]. Met behulp van een set van functies helpt DaQAPO met het identificeren van datakwaliteitsproblemen in de

event log. DaQAPO is volledig geïntegreerd met bupaR. BupaR (*Business Process Analytics in R*) is een open source R-pakket voor het uitvoeren van process mining in R [21].

Een andere tool om de datakwaliteit van event logs te boordelen is ProM. ProM (*Process Mining toolkit)* is een open source framework waar process mining algoritmen toegepast en nieuw geïmplementeerd kunnen worden [31]. Het implementeren van uitbreidingen gebeurd aan de hand van Java. In de meeste gevallen is de start input voor ProM een event log, met voorkeur in een XES-formaat [9]. Meerdere uitbreidingen zijn toegevoegd aan ProM om de datakwaliteit van event logs te beoordelen [9, 11, 15].

Nog een ander ontwikkelde tool om de datakwaliteit van event logs te beoordelen is QUELI. QUELI (*Querying Event Log for Imperfections*) is specifiek ontwikkeld om event log imperfecties op te sporen. Het uiteindelijke doel van QUELI is een tool te worden om een systematische, geautomatiseerde ondersteuning aan te bieden voor het opschonen van event logs [1]. De event log imperfecties die als inspiratie gebruikt zijn om deze tool te ontwikkelen, zijn die van Suriadi et al. [27]. Deze imperfecties worden in sectie 3.4.6 in meer detail besproken.

3.2 Datakwaliteitsverbetering

Nadat de datakwaliteit van de event log is geanalyseerd, zijn de datakwaliteitsproblemen in kaart gebracht [10]. Het verbeteren van de kwaliteit van event logs kan op twee manieren gebeuren: (1) het verbeteren van het vastleggen van de events terwijl de data worden gegenereerd of (2) het verbeteren van de data nadat ze zijn geregistreerd in een event log [4]. Bij het verbeteren van de event log zonder het registratieproces zelf aan te passen, is het mogelijk de event log op te schonen en bepaalde delen van de data te reconstrueren. Opschonen verwijst specifiek naar het identificeren van abnormale waarden in een dataset. Bij het reconstrueren van data worden bepaalde data vervangen of worden ontbrekende waarden aangevuld met betrouwbare alternatieven [23]. Het verbeteren van de event log kwaliteit kan louter op basis van de event log zelf of met behulp van externe bronnen die extra informatie bevatten [10].

In de literatuur zijn verschillende heuristieken opgesteld voor het verbeteren van de datakwaliteit. Deze heuristieken richten zich doorgaans op één specifiek datakwaliteitsprobleem en verschillende assumpties over de event log of de manier waarop het datakwaliteitsprobleem zich voordoet moeten voldaan zijn om deze toe te passen [10]. Er zijn methoden verzonnen om event log kwaliteit te verbeteren die zich specifiek focussen op het identificeren van afwijkende traces in een event log [5] en methoden die zich meer focussen op het reconstrueren van ontbrekende events in een event log [34]. Verder is er ook academisch werk dat zich focust op het opschonen en reconstrueren van event logs op het niveau van attribuut waarden [23].

3.3 Algemene taxonomieën

Hoewel een volledig overzicht van deze meer algemene taxonomieën buiten de scope van deze masterproef valt, zijn sommige concepten zeer relevant voor de process mining context [10]. Bij het identificeren van problemen met de datakwaliteit wordt vaak het begrip *fitness for use* gebruikt. Fitness for use zegt dat de kwaliteit van de data afhankelijk is van de mogelijkheid om de doelstellingen van de analyse correct te onderzoeken [35].

Verder bouwend op het fitness for use concept, definiëren Wang en Strong [35] een hiërarchisch raamwerk waarin ze 118 kenmerken voor datakwaliteit classificeren in twintig dimensies, die onderverdeeld worden in vier categorieën [35]. Rahm en Do [25] classificeren datakwaliteitsproblemen daarentegen op basis van het aantal verschillende data input bronnen. Voor zowel data uit één bron als data uit meerdere bronnen kunnen problemen op schema of instantie level voorkomen. Problemen op schema niveau komen voort uit problemen met het ontwerp van het data model design of een slechte toepassing van het invoeren van de data. Problemen op instantie niveau daarentegen zijn gerelateerd tot specifieke waarden van een dataveld [25].

Kim et al. [16] splitsen datakwaliteitsproblemen eerst op in ontbrekende en niet-ontbrekende data. De niet-ontbrekende data kunnen verder worden geclassificeerd in foute data en onbruikbare data [16]. Ook Müller en Freytag [22] categoriseren ontbrekende waarde-afwijkingen, maar definiëren de niet-ontbrekende data als syntactische en symantische waarde-afwijkingen [22]. Oliveira et al. [24] maken geen opsplitsing op basis van de aanwezigheid van de data, maar onderscheiden datakwaliteitsproblemen op basis van de granuliteit van de data. Gaande van het laagste granulariteitsniveau (attribuut/tupel) tot het hoogste (meerdere data bronnen) waar datakwaliteitsproblemen kunnen optreden [24].

Terwijl voorgaande taxonomieën eerder algemene datakwaliteitsproblemen bespreken, zijn er onderzoekers die op specifiek één type data focussen. Zo onderzocht Gschwandtner et al. [12] tijdsgerelateerde datakwaliteitsproblemen. De auteurs maken hierbij een onderscheid tussen data afkomstig uit één bron en data uit meerdere bronnen [12], zoals ook gebeurde in de taxonomie van Rahm en Do [25].

3.4 Taxonomieën voor process mining

Sectie 3.3 gaf een introductie tot enkele algemene taxonomieën van datakwaliteit. Hoewel er in deze taxonomieën ook datakwaliteitsproblemen besproken worden die eveneens relevant zijn voor event logs, houden ze geen rekening met de specifieke datastructuur van event logs. Event logs gaan gepaard met hun eigen specifieke datakwaliteitsproblemen, bijvoorbeeld gerelateerd aan de volgorde van events binnen een case [10]. Deze sectie zal daarom de belangrijkste taxonomieën van datakwaliteit op het gebied van

process mining in het algemeen bespreken. Voor een meer gedetailleerde uitleg wordt telkens verwezen naar bijbehorende tabellen die opgenomen zijn in de bijlagen in sectie 7.

3.4.1 Maturiteitsniveaus van het Process Mining Manifesto [30]

Een algemene taxonomie van event log datakwaliteit wordt gegeven in het Process Mining Manifesto. In deze taxonomie worden geen gedetailleerde datakwaliteitsproblemen gedefinieerd, maar wordt de datakwaliteit van event logs beoordeeld aan de hand van vijf maturiteitsniveaus [30]. Deze maturiteitsniveau's variëren van uitstekende kwaliteit (niveau vijf) tot slechte kwaliteit (niveau één) [10]. De specifieke uitleg van deze maturiteitsniveaus is opgenomen in Tabel 7.

Hoewel het mogelijk is om process mining technieken toe te passen op een event log van maturiteitsniveau één of twee, zijn de resultaten vaak zeer onbetrouwbaar. In de praktijk heeft het weinig nut om process mining toe te passen op logs van het laagste niveau. Om het nut van proces mining te maximaliseren, moeten bedrijven streven naar event logs van een zo hoog mogelijk maturiteitsniveau [30].

3.4.2 Datakwaliteitsproblemen van Bose et al. [6]

In tegenstelling tot het process mining manifesto [30] definieert Bose et al. [6] wel een set van specifieke datakwaliteitsproblemen. De 27 datakwaliteitsproblemen kunnen worden ondergebracht in vier categorieën [6]:

- **Ontbrekende data**: Verplichte data die ontbreken. Ontbrekende data zijn vaak het gevolg van problemen in het log proces.
- **Incorrecte data**: Beschikbare data die niet overeenkomen met wat gebeurd is in de realiteit.
- **Onnauwkeurige data**: De data zijn onnauwkeurig wanneer deze op een onvoldoende gedetailleerd niveau worden vastgelegd, waardoor de resultaten eventueel onbetrouwbaar worden.
- **Irrelevante data**: Irrelevant data zijn data die eerst filtering of transformatie vereisen voordat deze gebruikt kunnen worden in de proces analyse.

De event log datakwaliteitsproblemen kunnen geplaatst worden in een matrix als een combinatie van enerzijds de bovenstaande categorieën en anderzijds de volgende negen event log componenten: case, event, relatie, case attribuut, positionering, activiteit naam, tijdstip, resource, en event attribuut [6]. Een beschrijving van alle 27 event log datakwaliteitsproblemen zijn opgenomen in Tabel 8.

3.4.3 Datakwaliteitsdimensies van Verhulst [9]

De taxonomie van Verhulst [9] is gebaseerd op zowel algemene literatuur over datakwaliteit als de meer specifieke event log literatuur zoals de taxonomie van Bose et al. [6] en de aanpak om datasets te transformeren naar event logs van onderzoeker van der Aalst [29]. Op basis van een analyse van deze literatuur worden twaalf datakwaliteitsdimensies voor event log data gespecificeerd [9]. De twaalf dimensies worden besproken in Tabel 9.

Verhulst [9] definieert een meetmethode om de volgende dimensies te kwantificeren: volledigheid, uniekheid, tijdigheid, geldigheid, nauwkeurigheid, consistentie, relevantie, vertrouwelijkheid, en formaat. Elke dimensie krijgt een score op tien of een *boolean* waar/onwaar oordeel om zo de event log te beoordelen. Voor sommige dimensies wordt daarentegen gesteld dat er geen meting mogelijk is zonder extra inbreng, bijvoorbeeld het oordeel van een gebruiker. Bij bijvoorbeeld de geloofwaardigheid dimensie moet de gebruiker de geloofwaardigheid van de event log beoordelen op basis van de eigen interpretatie [9].

3.4.4 Datakwaliteitsdimensies van Kherbouche et al. [15]

De taxonomie van Kherbouche et al. [15] beoordeelt de kwaliteit van de event log op basis van vier datakwaliteitsdimensies, die telkens worden beoordeeld met behulp van een set metrieken. De metrieken worden per datakwaliteitsdimensie verder besproken in Tabel 11. De vier datakwaliteitsdimensies zijn [15]:

- **Complexiteit**: Deze dimensie beschrijft de begrijpelijkheid van het model indien de event log wordt omgezet tot een procesmodel. Er is een duidelijke correlatie tussen de complexiteit van event logs en de begrijpelijkheid van de ontdekte procesmodellen.
- **Nauwkeurigheid**: De event log data zijn representatief voor wat er zich in de praktijk heeft voorgedaan. Onnauwkeurige event log data belemmeren mogelijks de ontdekking van zinvolle analyseresultaten en inzichten.
- **Consistentie**: Deze dimensie probeert de ruis, het aantal onlogische traces en andere gelijksoortige karakteristieken in de event log in te schatten.
- **Volledigheid**: Is alle vereiste informatie aanwezig en voldoende gedetailleerd is om de gestelde process mining vragen te beantwoorden. Niet alle beschikbare data zijn relevant om bepaalde inzichten te verkrijgen, maar wanneer de aanwezigheid van deze data vereist is voor bepaalde types van process mining wordt het ontbreken ervan problematisch.

3.4.5 Conceptueel datakwaliteit kader van Lu en Fahland [18]

Lu en Fahland [18] stellen een conceptueel kader voor dat helpt begrijpen hoe datakwaliteitsproblemen kunnen voorkomen en met elkaar gelinkt kunnen worden. Het conceptuele kader wordt gevisualiseerd als een tabel. De kolommen vertegenwoordigen de entiteiten van input data, namelijk de kwaliteit van events, kwaliteit van de volgorde van events of relaties tussen events, en kwaliteit van labels van events. De rijen vermelden twee dimensies van datakwaliteit die de kwaliteit van de event beoordelen, elk vanuit hun eigen perspectief. Deze dimensies zijn individuele betrouwbaarheid en globale zekerheid. Een goede event log moet zowel individueel betrouwbaar zijn als globale zekerheid hebben om analyses op uit te voeren [18]. Het conceptuele kader wordt visueel voorgesteld in Tabel 3.

Dimensies van datakwaliteit	**Kwaliteit van events**	**Kwaliteit van de volgorde van events of relaties tussen events**	**Kwaliteit van labels van de events**
Individuele betrouwbaarheid	+: 100% betrouwbaar -: grotendeels betrouwbaar - -: grotendeels onbetrouwbaar		
Globale zekerheid	+: pad, structuur of patroon is significant -: pad, structuur of patroon is niet significant genoeg		

Tabel 3: Conceptueel kader van Lu en Fahland [18]

Individuele betrouwbaarheid beoordeelt de event data vanuit een data perspectief. Meer specifiek verwijst de individuele betrouwbaarheid naar de mate waarin de gebruiker gelooft dat de event data de realiteit weerspiegelt. Er zijn drie mogelijke waarden om de individuele betrouwbaarheid uit te drukken: (1) '+', wat 100% betrouwbaarheid weerspiegelt, (2) '-', de data zijn grotendeels betrouwbaar, maar er is een minderheid van onbetrouwbaarheid aanwezig, en (3) '- -' wat wilt zeggen dat de gebruiker de data voor de grote meerderheid als onbetrouwbaar beschouwd [18].

Globale zekerheid beoordeelt event data vanuit een analyse perspectief. De globale zekerheid bespreekt of een aanwezig pad, structuur of patroon niet enkel willekeurig of uniek voorkomt in de event log. De gebruiker kan hier twee mogelijke waarden invullen: (1) '+' wanneer het pad, de structuur of het patroon significant genoeg is om te worden waargenomen, en (2) '-' in het geval het waargenomen mechanisme niet significant genoeg is [18].

3.4.6 Event log imperfectie patronen van Suriadi et al. [27]

Suriadi et al. [27] definiëren elf imperfectie patronen op basis van eigen ervaring met real-live cases, specifiek voor event logs. Omdat er patronen worden gebruikt om datakwaliteitsproblemen te identificeren, is het mogelijk om oplossingen voor deze terugkerende problemen te ontwikkelen die telkens toegepast kunnen worden indien zo een imperfectie patroon zich voordoet [27]. Deze elf geïdentificeerde imper-

fectie patronen zijn uitgelegd in Tabel 10. De event log imperfecties door Suriadi et al. [27] zijn gebruikt als input voor het ontwikkelen van QUELI (eerder besproken in sectie 3.1). Meer specifiek kan QUELI (bij de tijd van schrijven) vijf van de elf event log imperfecties detecteren: op formulieren gebaseerde event registratie, onbedoeld tijdreizen, *collateral events*, label synoniemen, en gelijknamige labels [1].

3.4.7 Indicatoren voor event volgorde problemen van Dixit et al. [8]

Dixit et al. [8] definiëren drie klassen van indicatoren die helpen met het lokaliseren van problemen van de volgorde van events. Deze indicatoren zijn gebaseerd op basis van de reeds bestaande literatuur over datakwaliteitsproblemen in event logs [6, 19, 27, 30] en over tijdstip georiënteerde problemen bij data in het algemeen [12]. Indien een indicator aanwezig is, wil dit niet met zekerheid zeggen dat er een probleem is met de volgorde van events in de event log. Het doel van de indicatoren is de analist te ondersteunen bij het selecteren van de best mogelijke opschoon acties. De drie gedefinieerde indicatoren zijn granuliteit, volgorde onregelmatigheden, en statistische onregelmatigheden [8].

Granulariteit is een indicator van problemen met de volgorde van events indien de tijdstippen niet gedetailleerd genoeg gelogd zijn of indien er een gemengde set van tijdstippen granulariteit aanwezig is in de event log. De granulariteit van een tijdstip is afhankelijk van hoe gedetailleerd een tijdstip gelogd is (tot op de seconde, milliseconde, ...). Indien de granulariteit van tijdstippen niet gedetailleerd genoeg is, kunnen tijdstippen onnauwkeurig gelogd worden. Verder kan ook een combinatie van verschillende niveaus in granulariteit van tijdstippen leiden tot een foutieve ordening van events. Bijvoorbeeld in een situatie waarbij een eerste event vastgelegd wordt tot op de seconde (bijvoorbeeld 05/06/2022 13:24:53) en een hierop volgende event slechts tot op de dag (bijvoorbeeld 05/06/2022 00:00:00). Het verschil in de granulariteit van de tijdstippen zorgt ervoor dat deze events omgekeerd voorkomen in de event log dan dat deze zich in de realiteit hebben afgespeeld [8].

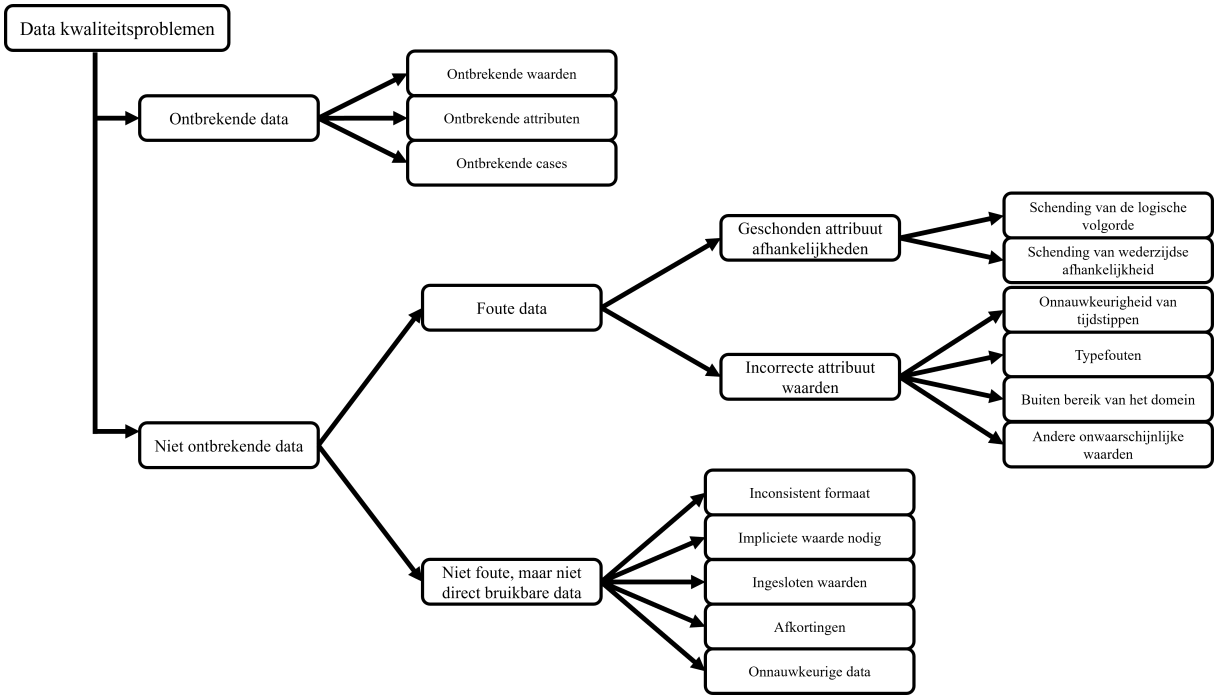
Volgorde imperfecties kunnen worden gelokaliseerd door ongebruikelijke ordening te identificeren. Ook ontbrekende events en incorrecte tijdstippen kunnen worden geïdentificeerd door te kijken of er ongebruikelijke ordeningen aanwezig zijn tussen activiteiten. Ontbrekende events en incorrecte tijdstippen zijn het gevolg van events die pas achteraf zijn vastgelegd of events die manueel zijn ingegeven [8].

Ten slotte kunnen ook **statistische imperfecties** wijzen op het bestaan van tijdstip gerelateerde problemen. Een voorbeeld hiervan is het identificeren van de relatie tussen twee activiteiten per case. Een ander voorbeeld is de variatie van tijdstip formaten van de events. Als een event log is opgebouwd met events van verschillende systemen, kunnen tijdstippen in verschillende formaten vastgelegd worden, waardoor deze eventueel foutief geïnterpreteerd kunnen worden. Bijvoorbeeld, indien in de event log voor een

bepaalde activiteit enkel maar tijdstippen voorkomen met een datumwaarde tussen de één en twaalf en geen enkele datumwaarde groter dan twaalf, is de kans groot dat de tijdstippen van deze activiteit gelogd worden volgens het MM/DD/YYYY formaat in plaats van DD/MM/YYYY formaat. Verder zijn er ook andere statistische imperfecties om problemen met de event volgorde te ontdekken, bijvoorbeeld het gebruik van batchverwerking en het probleem met meerdere tijdzones [8].

3.4.8 Datakwaliteitsproblemen van Vanbrabant et al. [33]

Om een gestructureerd inzicht te geven over de diverse datakwaliteitsproblemen van event logs, stellen Vanbrabant et al. [33] een gesynthetiseerde taxonomie van zowel algemene als zorgspecifieke taxonomieën voor [10]. De taxonomie bevat een set van generieke categorieën en sub-categorieën, die consistent zijn met die van Kim et al. [16]. In deze categorieën worden veertien specifieke datakwaliteitsproblemen van event logs ingedeeld [33]. De veertien kwaliteitsproblemen worden in Tabel 12 besproken en zijn samen met de gehele taxonomie visueel voorgesteld in Figuur 2.



Figuur 2: Visuele voorstelling taxonomie van Vanbrabant et al. [33]

Vanbrabant et al. [33] maken eerst een onderscheid tussen ontbrekende en niet ontbrekende data. Ontbrekende data hebben betrekking op data die niet in de event log opgenomen zijn, terwijl dit wel het geval zou moeten zijn. De categorie niet-ontbrekende data daarentegen bevat problemen met de kwaliteit van de waarden die in de event log zijn vastgelegd. Binnen deze categorie wordt onderscheid gemaakt tussen foute data enerzijds en niet foute, maar niet direct bruikbare data anderzijds. De eerste subcategorie, foute data, kan verder worden verdeeld in de volgende sub-subcategorieën: geschonden attribuut afhankelijkheden en incorrecte attribuutwaarden. Voor de tweede subcategorie, niet foute, maar niet di-

rect bruikbare data, zijn eerst transformaties of een andere bewerkingen vereist om deze voor analyse doeleinden te kunnen gebruiken [33].

3.4.9 Vijftien datakwaliteitsmetrieken van Fischer et al. [11]

Dimensie	**Synoniemen**	**Kwantificeerbaar**
Beschikbaarheid	Toegankelijkheid	Nee
Nauwkeurigheid	Correctheid, exactheid, feit, foutenvrij, granulariteit, expertiseniveau, ruis, precisie, synchroniciteit, variantie	Ja
Hoeveelheid data	Inhoud, dekking, hoeveelheid	Nee
Geloofwaardigheid	Vertrouwelijkheid, *credibility*, *faithfullness*, plausibiliteit, redelijkheid, betrouwbaarheid, toereikendheid, vertrouwen, waarheidsgetrouwheid	Nee
Volledigheid	Gegevensverval, *disaggregation*	Ja
Beknoptheid		Ja
Consistentie	Rangschikking, samenhang, *cohesiveness*, compatibiliteit, vergelijkbaarheid, formaliteit, formaat, integriteit, draagbaarheid, structuur, geldigheid	Ja
Data management	Onderhoudbaarheid	Nee
Interlinking		Nee
Interpreteerbaarheid	Compliance, conformiteit, gegevensspecificatie, metadata, natuurlijkheid, semantische stabiliteit, gebruik van standaarden	Ja
Objectiviteit	Onbevooroordeeld	Nee
Presentatie kwaliteit	Uiterlijk, aantrekkelijkheid, navigeerbaarheid, reactie tijd, latentie	Nee
Prijs		Nee
Relevantie	*Fitness*, belang, neutraliteit, redundantie, weerspiegeling van de werkelijkheid, representativiteit	Nee
Reputatie		Nee
Beveiliging		Nee
Service kwaliteit		Nee
Specialisatie		Nee
Tijdigheid	*Currency*, vervaldatum, versheid, periodiciteit, volatiliteit	Ja
Traceerbaarheid	Controleerbaarheid, documentatie, manipulatiegemak, bestaan, herkomst, herhaalbaarheid, herstelbaarheid, verhandelbaarheid, transparantie	Ja
Begrijpelijkheid	Duidelijkheid, complexiteit, begrijpelijkheid, bedieningsgemak, gemak om te begrijpen, gebruiksgemak, leesbaarheid	Nee
Uniekheid	Duplicaten	Ja
Bruikbaarheid	Klantenondersteuning, effectiviteit, efficiëntie, prestaties, nut, *utility*	Nee
Waarde-toevoegend		Nee
Verifieerbaarheid		Nee

Tabel 4: Dimensies van Fischer et al. [11]

Op basis van besproken datakwaliteitsdimensies in 48 papers over datakwaliteit hebben Fischer et al. [11] een lijst met 118 datakwaliteitsdimensies geclusterd tot een set van 25 verschillende datakwaliteitsdimensies. Per dimensie zijn de verschillende synoniemen vermeld alsook of die specifieke dimensie kwantificeerbaar is zonder menselijke input [11]. De lijst met verschillende datakwaliteitsdimensies is opgenomen in Tabel 4.

Om een zo automatisch mogelijke tool te ontwikkelen hebben Fischer et al. [11] zich eerst beperkt tot de acht datakwaliteitdimensies die kwantificeerbaar zijn zonder de nood aan menselijke input. Verder hebben ze nog vier datakwaliteitsdimensies uitgesloten die focussen op labeling omdat de intentie was om de kwaliteit van numerieke waarden (zoals tijdstippen) te beoordelen. Dit heeft geleid tot de volgende set van vier datakwaliteitsdimensies [11]:

- **Nauwkeurigheid** verwijst naar het verschil tussen de waarde weergegeven in de event log en de waarde die zich in de realiteit heeft voorgedaan. Elke metriek die onnauwkeurige tijdstippen onderzoekt, is toegewezen aan deze dimensie.
- **Volledigheid** kijkt of voor één specifieke variabele alle waardes opgenomen zijn in de event log.
- **Consistentie** betekent dat de data waarden in alle events op een gelijkwaardige manier moeten worden weergegeven. Aangezien er meerdere eenheden nodig zijn om consistentie te evalueren, is het onmogelijk om de consistentie voor één enkel event op event niveau te beoordelen.
- **Uniekheid** verwijst naar ongewenste duplicaten binnen of tussen systemen voor een bepaalde event log. Net zoals bij de dimensie consistentie, zijn er meerdere eenheden nodig om ongewenste duplicaten te hebben. Hierdoor is het onmogelijk om de uniekheid voor één enkel event op event niveau te beoordelen.

De vier datakwaliteitsdimensies worden gekwantificeerd aan de hand van vijftien metrieken op event log niveau, trace niveau, activiteit niveau en event niveau [11]. Deze metrieken worden in meer detail besproken in Tabel 13.

3.4.10 Recapitulatie taxonomieën

Zoals besproken in deze sectie, hebben process mining onderzoekers verschillende taxonomieën en concepten ontwikkeld om datakwaliteitsproblemen in event logs te identificeren om zo de betrouwbaarheid van analyseresultaten te verbeteren [36]. Het Process Mining Manifesto definieert vijf maturiteitsniveaus van event logs, variërend van slechte kwaliteit tot een uitstekende kwaliteit. Deze maturiteitsniveaus geven een duidelijke indicatie in hoeverre er vertrouwd kan worden op de validiteit van de event log data [30], maar geeft geen concrete richtlijnen om het correcte maturiteitsniveau te bepalen [15]. Een meer specifieke taxonomie om datakwaliteitsproblemen van event logs te identificeren is voorgesteld door

Bose et al. [6]. Bose et al. identificeren vier brede categorieën die zich voordoen bij het analyseren van de kwaliteit van event logs in process mining, namelijk ontbrekende data, onjuiste data, onnauwkeurige data en irrelevante data. Verder worden 27 specifieke datakwaliteitsproblemen in deze vier categorieën geclassificeerd [6]. Bose et al. hebben andere process mining onderzoekers sterk aangemoedigd om zich meer te cocnentreren op technieken die deze datakwaliteitsproblemen aanpakken [20]. Eén van deze onderzoekers is Verhulst [9], die op basis van onder andere Bose et al. [6] een raamwerk heeft gecreëerd om de kwaliteit van event data te meten. Twaalf dimensies om datakwaliteitsproblemen in event logs te kwantificeren worden in dit raamwerk gedefinieerd [9].

Op basis van de literatuur werden door Kherbouche et al. [15] vier dimensies voor datakwaliteitsproblemen in event logs aangevuld met meerdere metrieken per dimensie om deze te kwantificeren [15]. Lu en Fahland [18] focussen niet specifiek op datakwaliteitsproblemen, maar hebben een conceptueel kader ontwikkeld om datakwaliteitproblemen beter te kunnen presenteren en met elkaar te linken. In dit kader wordt de kwaliteit van events, volgorde van events of relaties tussen events, en labels van events beoordeelt via zowel een individuele betrouwbaarheid als een globale zekerheid perspectief [18].

Gebaseerd op hun eigen ervaringen bij het omzetten van rauwe data in event logs, hebben Suriadi et al. [27] elf event log imperfectie patronen gedefinieerd [27]. Aanvullend op deze imperfectie patronen [27], alsook op het Process Mining Manifesto [30], Bose et al. [6], en de meer specifieke literatuur over tijdstip georiënteerde data problemen van Gschwandtner et al. [12] hebben Dixit et al. [8] drie indicatoren gedefinieerd om event volgorde problemen te detecteren, namelijk granulariteit, volgorde imperfecties, en statistische imperfecties. Alle drie de indicatoren zijn ook duidelijk geoperationaliseerd om deze indicatoren te kunnen opsporen [8].

Vanbrabant et al. [33] creëren een taxonomie met datakwaliteitsproblemen gebaseerd op de eerder bestaande literatuur [6, 12, 14, 16, 19, 22, 24, 25, 35]. Een lijst van veertien datakwaliteitsproblemen worden verder onderverdeeld en opgesplitst in verschillende categorieën [33]. Ten slotte clusteren Fischer et al. [11] op basis van 48 papers over datakwaliteit een lijst van 118 datakwaliteitsdimensies tot een set van 25 verschillende datakwaliteitsdimensies. Uit de acht dimensies die kwantificeerbaar zijn, selecteren Fischer et al. [11] vier datakwaliteitdimensies om metrieken te ontwikkelen die de datakwaliteit beoordelen [11].

4 Benchmarking studie

In deze sectie wordt de benchmarking studie uitgevoerd op enkele publieke real-life event logs. Eerst worden de gebruikte tools om datakwaliteit te beoordelen besproken in subsectie 4.1. Verder wordt verklaard waarom de geselecteerde event logs gebruikt zijn en worden deze kort beschreven in subsectie

4.2. De gebruikte datakwaliteitsdimensies alsook de metrieken om de geselecteerde dimensies te kwantificeren worden verder besproken in subsectie 4.3. Ten slotte wordt de benchmarking studie uitgevoerd en worden de resultaten gepresenteerd in subsectie 4.4.

4.1 Tools

De datakwaliteit van de event logs werd beoordeeld met behulp van beschikbare tools over de besproken taxonomieën (sectie 3.4). Verder worden er ook aanvullende analyses uitgevoerd om een overzicht te verkrijgen van de geselecteerde event logs. Er worden analyses uitgevoerd in de tools R, Apromore, en ProM. Om een algemeen overzicht te krijgen van de event log statistieken, wordt het R-pakket bupaR gebruikt. BupaR is een verzameling R-pakketten die een framework bieden voor reproduceerbare procesanalyse in R. Eén van de functionaliteiten van bupaR is het uitvoeren van exploratieve en beschrijvende event log analyses [13]. Verder wordt het R-pakket DaQAPO gebruikt om de kwaliteit van event logs te beoordelen. DaQAPO biedt functies aan die het mogelijk maken om datakwaliteitsproblemen van event logs te identificeren. DaQAPO is volledig geïntegreerd met bupaR, een open source R-pakket voor het uitvoeren van process minig in R [21].

Apromore is een ondersteunde open-source tool voor Process Mining[1]. In Apromore werd voor elke event log een standaard BPMN model ontwikkeld om het basis proces te visualiseren. Een BPMN, ofwel *Business Process Model and Notation*, toont op een grafische manier de opeenvolging van de activiteiten in een procesmodel. Het primaire doel van BPMN is het bieden van een notatie die begrijpbaar is voor een breed spectrum aan gebruikers om processen te visualiseren in verschillende bedrijfscontexten [7].

In ProM zullen de de beschikbaar gestelde extensies van Fischer et al. [11] en Verhulst [9] gebruikt worden om event logs te beoordelen aan de hand van verschillende metrieken per datakwaliteitdimensie. Deze datakwaliteitdimensies en metrieken worden in meer detail besproken in sectie 4.3.

4.2 Event log selectie

De gekozen event logs zijn event logs van de jaarlijkse BPI challenge. Er is gekozen om de event logs te nemen van de meest recente BPI challenges. De geselecteerde event log zijn: BPI challenge 2017, BPI challenge 2018, BPI challenge 2019, BPI challenge 2020 domestic declaration, en BPI challenge 2020 request for payment. In de rest van deze sectie wordt respectievelijk verwezen naar deze event logs als 2017, 2018, 2019, 2020-DD, en 2020-RFP. Alle event logs zijn afkomstig van de website van 4TU.ResearchData, *an international data repository for science, engineering and design*[2].

[1] https://apromore.com/
[2] https://data.4tu.nl/

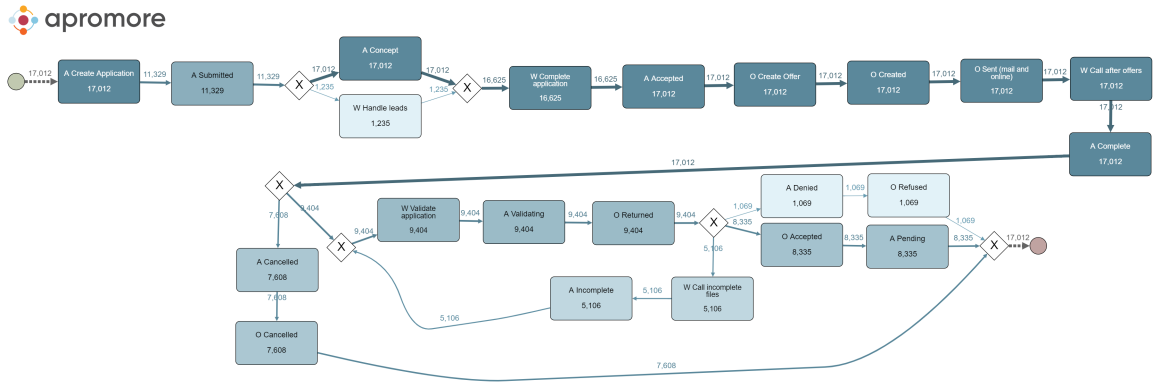
Aan de hand van bupaR werd een duidelijk overzicht met enkele basis statistieken per event log verkregen. Om deze event logs in bupaR in te laden, was het toepassen van preprocessing stappen vereist. Deze zullen specifiek per event log overlopen worden wanneer elke event log in meer detail besproken wordt. Verder werd, met behulp van de tool Apromore, voor elke event log een standaard BPMN model ontwikkeld om het basis proces te visualiseren. Het procesmodel werd gecreëerd op basis van een versimpelde set van traces in de event log. Niet alle mogelijke traces zijn getoond in de BPMN's omdat dit zou resulteren in een te groot model dat geen overzicht meer biedt over het procesverloop. Algemene informatie over de verschillende event logs is opgenomen in Tabel 5. Er zijn drie relatief grote event logs (2017, 2018, 2019) opgenomen en twee kleinere event logs (2020-DD en 2020-RFP).

	2017	**2018**	**2019**	**2020-DD**	**2020-RFP**
#Events	1 160 405	2 396 132	1 595 923	56 437	36 796
#Attributen	19	75	22	11	16
#Cases	31 509	43 809	251 734	10 500	6 886
#Traces	10 452	34 496	11 973	99	89
#Cases/#Trace	3,015	1,27	21,03	106,06	77,37
#Unieke traces	2 924 (28%)	26 602 (77%)	9 030 (75%)	46 (46%)	41 (46%)
#Verschillende activiteiten	26	39	42	17	19
Minimum trace lengte	8	22	1	1	1
Mediane trace lengte	22	50	11	8	7
Maximum trace lengte	61	2971	990	24	20
Gemiddelde case duur	3,12 weken	11,02 maanden	2,28 maanden	1,65 weken	1,72 weken

Tabel 5: event log basisinformatie overzicht

4.2.1 BPI challenge 2017

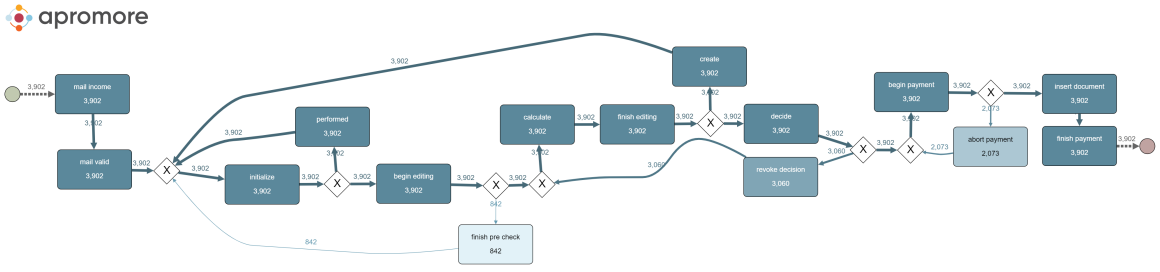
De event log van de BPI challenge 2017 bevat events van een lening aanvraagproces van een Nederlandse financiële instelling. Alle aanvragen die vanaf januari 2016 tot februari 2017 zijn ingediend via een online systeem zijn opgenomen in de event log. Om de event log te gebruiken was geen preprocessing vereist. De dataset kon zonder verdere manipulaties van de data omgezet worden in een event log. De event log bevat events van 31 509 lening aanvragen. Op basis van de algemene statistieken zijn geen uitzonderlijke uitschieters ten opzichte van de andere event logs waargenomen en hierdoor wordt deze event log gezien als een goed algemeen referentiepunt. Bovendien is dit de event log met relatief gezien het minste unieke traces, ofwel traces die maar één case representeren. Het BPMN model van deze event log is voorgesteld in Figuur 3.



Figuur 3: BPMN van het BPI challenge 2017 proces

4.2.2 BPI challenge 2018

De event log van de BPI challenge 2018 bevat events over aanvragen voor rechtstreekse betalingen door het Europees Landbouwgarantiefonds aan Duitse boeren. Om de dataset in R om te zetten tot een event log, zijn enkele preprocessing stappen uitgevoerd. Zo werd in de event log de string "0;n/a"gebruikt om naar missing resources te verwijzen. In totaal bevat de event log events over 43 809 aanvragen over een periode van drie jaar. Deze event log bevat het meeste unieke traces. De kortste trace bevat 24 events, de langste 2 971 en de mediane waarde is 50 events per case die opgebouwd zijn uit een set van 14 verschillende activiteiten. De event log bevat veel variatie in de lengte van de traces. Bovendien zal de duurtijd per case, vergeleken met de andere geselecteerde event logs, gemiddeld lang zijn, namelijk 11,02 maanden. Daarenboven is dit ook de grootste geselecteerde event log die op basis van de informatie uit Tabel 5 ook de meest complexe van de selectie lijkt. Het procesmodel voor het aanvraagproces is opgenomen in Figuur 4.

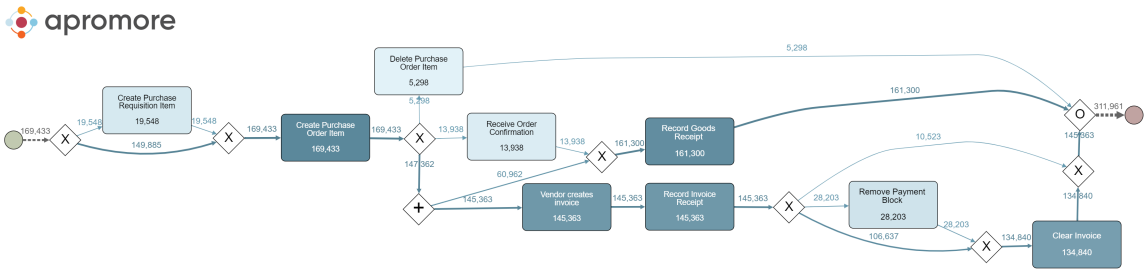


Figuur 4: BPMN van het BPI challenge 2018 proces

4.2.3 BPI challenge 2019

De event log van de BPI challenge 2019 bevat events over het inkooporder proces. In de event log zijn de ontbrekende waarden correct geregistreerd. Meer dan 1,5 miljoen events zijn verzameld over

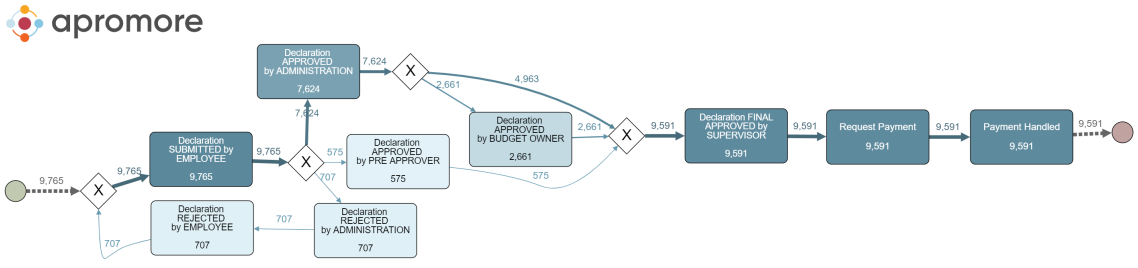
inkooporders die in 2018 zijn ingediend. De data verwijzen naar het inkoop-tot-betaalproces (zonder de goedkeuringsworkflow van de inkooporders en de facturen) voor verschillende categorieën goederen en diensten en omvatten veel verschillende soorten leveranciers. Er zijn in in totaal 251 734 cases. In deze cases zijn er 1 595 923 events met betrekking tot 42 activiteiten. In vergelijking met de event logs van 2017 en 2018 heeft deze event log relatief weinig traces (vergeleken met het aantal cases) en is de gemiddelde lengte van de trace relatief kort, namelijk zes events per case. Desondanks zijn er traces opgenomen in de event log met een lengte van 990. Deze grote variatie in trace lengte is één van de factoren die deze event log zeer interessant maakt om op te nemen in deze benchmarking studie. Het procesmodel is opgenomen in Figuur 5.



Figuur 5: BPMN van het BPI challenge 2019 proces

4.2.4 BPI challenge 2020

Ten slotte zijn er twee event logs opgenomen van de BPI challenge van 2020. Deze event logs beschrijven het proces om claims (om reiskosten terug te vorderen) te verwerken. De geselecteerde event logs bevatten data over claims van binnenlandse aangiften (2020-DD) en claims over uitgaven die niet reis gerelateerd zijn (2020-RFP).

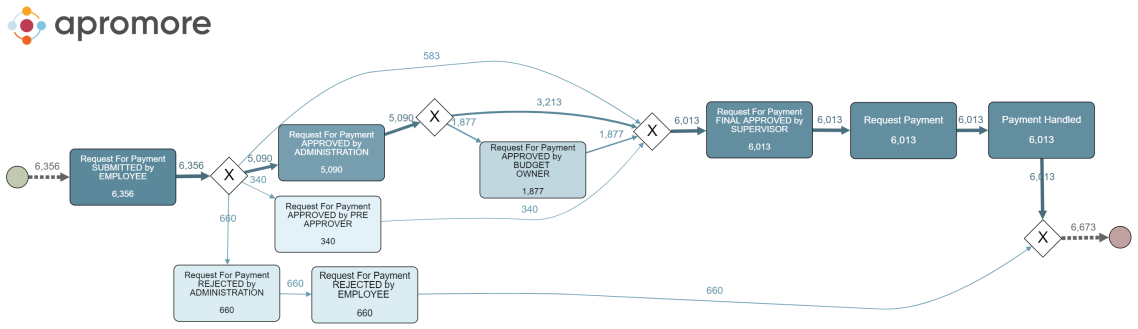


Figuur 6: BPMN van het BPI challenge 2020 Domestic Declarations proces

De **Domestic Declaration (DD)** event log bevat 56 437 events over 10 500 cases. Dit is de event log waarbij relatief gezien (ten opzichte van het aantal cases) het minste traces aanwezig zijn in de event log. Ook is dit de event log met het minst aantal attributen. Hierdoor lijkt op basis van de algemene

informatie dat deze event log de minst complexe event log van de selectie is. Het procesmodel van dit proces is visueel voorgesteld in Figuur 6.

De event log **Request For Payment (RFP)** bevat 6 886 cases over 36 796 events. Deze event log bevat geen starttijd per event. Dit is de kleinste event log in aantal cases met het minst aantal verschillende traces, toch zijn er meer verschillende activiteiten dan de 2020-DD event log. Ook vertegenwoordigt elke trace minder cases en bevat de event log meer attributen, wat erop lijkt te wijzen dat deze event log complexer is dan de 2020-DD event log. Het procesmodel voor deze event log is visueel opgenomen in Figuur 7.



Figuur 7: BPMN van het BPI challenge 2020 Request for payment proces

4.3 Evaluatie dimensies en metrieken

Tabel 14, opgenomen in de bijlage (sectie 7) start vanuit de datakwaliteitsdimensies gedefinieerd door Fischer et al. [11]. De datakwaliteitsproblemen en imperfectie patronen, gezien in de andere taxonomieën, worden in de 25 dimensies van Fischer et al. [11] geclassificeerd. Tabel 14 zal de basis vormen voor de te benchmarken metrieken. Verder wordt per dimensie gekeken of deze kwantificeerbaar is op basis van enkel de event log.

Omdat er in deze masterproef geen toegang is tot een domeinexpert en extra informatie over het registreren van event data, zal enkel gefocust worden op datakwaliteitsdimensies die kwantificeerbaar zijn op basis van louter de event log: nauwkeurigheid, volledigheid, consistentie, en uniekheid. Deze datakwaliteitdimensies zijn gekwantificeerd met behulp van metrieken die de datakwaliteit van de event log voor die dimensie beoordelen. Omdat meerdere tools gebruikt zijn om de event log kwaliteit te beoordelen, zijn er ook meerdere metrieken gebruikt om eenzelfde datakwaliteitsprobleem te kwantificeren. Dit is bijvoorbeeld het geval bij de metrieken granulariteit door Fischer et al. [11] en tijdstip formaat door Verhulst [9]. Beide metrieken beoordelen hoe specifiek de tijdstippen in de event log zijn geregistreerd. Het is echter interessant om beiden te behouden om af te leiden of beide metrieken de kwaliteit van dit datakwaliteitsprobleem gelijkaardig kwantificeren. De gebruikte dimensies en metrieken in de tool ProM

zijn samengevat per auteur in Tabel 15 in de bijlagen (sectie 7).

4.4 Resultaten

Datakwaliteitsdimensies en metrieken	2017	2018	2019	2020-DD	2020-RFP
Algemene gemiddelde datakwaliteit	89%	73%	80%	88%	87%
Nauwkeurigheid	79%	78%	82%	94%	92%
Infrequente volgorde van activiteiten	21%	0%	65%	99%	100%
Overlappende activiteiten per resource	87%	100%	100%	100%	100%
Overlappende activiteit	*95%*	*100%*	*100%*	*NA*	*NA*
Toekomstige invoer	100%	100%	100%	100%	100%
Tijdstip correctheid	*100%*	*100%*	*100%*	*100%*	*100%*
Granulariteit	100%	89%	43%	69%	69%
Tijdstip formaat	*100%*	*100%*	*80%*	*90%*	*90%*
Negatieve duur van activiteiten	90%	100%	100%	100%	NA
Volledigheid	77%	85%	85%	85%	82%
Ontbrekende trace	99%	100%	100%	98%	98%
Ontbrekende activiteit	100%	100%	100%	100%	100%
Ontbrekend event	1%	0%	0%	0%	0%
Ontbrekend tijdstip	100%	100%	100%	100%	100%
Event volgorde per trace	*100%*	*100%*	*100%*	*100%*	*100%*
Ontbrekende waarden	90%	90%	90%	90%	100%
Event-Resource controle	100%	100%	100%	100%	100%
Transactionele informatie controle	100%	90%	90%	90%	60%
Attribuut relevantie controle	30%	100%	100%	100%	100%
Consistentie	98%	75%	92%	92%	92%
Formaat	100%	100%	100%	100%	100%
Gemengde granulariteit van de log	100%	58%	95%	99%	100%
Gemengde granulariteit van traces	100%	59%	90%	94%	94%
Gemengde granulariteit van activiteiten	100%	71%	98%	99%	100%
XES Standaarden	100%	100%	100%	100%	100%
Consistentie controle	90%	60%	70%	60%	60%
Uniekheid	100%	56%	61%	80%	81%
Duplicaten in de event log	100%	43%	6%	41%	44%
Duplicaten in de event log	*100%*	*100%*	*10%*	*100%*	*100%*
Duplicaten in een trace	100%	43%	83%	100%	100%
Duplicaten in een trace	*100%*	*90%*	*90%*	*100%*	*100%*
Duplicaten in een activiteit	100%	81%	92%	100%	100%

Tabel 6: Resultaten datakwaliteitsbeoordeling van geselecteerde event logs. Indien meerdere metrieken eenzelfde datakwaliteitsprobleem kwantificeren, is in het grijs de metriek opgenomen die de andere metriek ondersteund.

In Tabel 6 is per datakwaliteitsdimensie en metriek de gekwantificeerde score weergegeven. Per metriek werd door de analyses van Fischer et al. [11], Verhulst [9], of de DaQAPO functies [21] een score gevonden. De scores van Fischer et al. [11] zijn telkens uitgedrukt in percentages terwijl de scores van Verhulst [9] uitgedrukt worden als een score op 10. De score op 10 is in de tabel omgezet naar percentages. De DaQAPO functies resulteren in een percentage slechte events. Om dit percentage om te

zetten naar een gekwantificeerde score, wordt dezelfde techniek als Verhulst [9] gebruikt om bijvoorbeeld een score te geven aan de ontbrekende waarden. Iedere metriek start met een score van 100%, per 1% slechte cases wordt er telkens een straf van 10% afgetrokken van de score. Zo zullen 5% ontbrekende waarden resulteren in een score van 50%.

Indien bepaalde metrieken hetzelfde datakwaliteitsprobleem kwantificeren, is de methode van Fischer et al. [11] gebruikt om dit probleem te kwantificeren. De metrieken door Fischer et al. [11] zijn ontwikkelt in recenter onderzoek en is complexer en statistisch correcter berekend dan die van Verhulst [9]. In de tabel zijn echter beide metrieken opgenomen. De metriek die niet is gebruikt om de score voor de bovenliggende dimensie te berekenen, is opgenomen in een grijze kleur. Deze extra waarde kan interessant zijn om gelijkenissen en eventuele tegensprekende resultaten tussen beide methodes te ontdekken. Het gemiddelde van de scores van de gebruikte metrieken werd berekend om een score toe te kennen aan de bijbehorende datakwaliteitsdimensies, net zoals Fischer et al. [11] doet in de door hun gemaakte ProM extensie. Hierbij werd de assumptie gemaakt dat elke metriek even belangrijk is om de datakwaliteit van de bijhorende dimensie te kwantificeren. Ten slotte is het gemiddelde van de datakwaliteitsdimensies berekend om een algemene gemiddelde datakwaliteitsscore toe te kennen aan elke event log. Ook hier is de assumptie gemaakt dat elke dimensie even belangrijk is om de datakwaliteit van een event log te beoordelen.

In eerste instantie is af te leiden dat de datakwaliteit van de 2017 event log de beste algemene datakwaliteit heeft. Verder zullen de twee kleinere event logs uit de BPI challenge van 2020 ook een hoge algemene datakwaliteit hebben, terwijl deze van 2018 duidelijk de slechtste datakwaliteit heeft.

Bij een meer gedetailleerde analyse van de datakwaliteitsdimensies en metrieken, valt op dat de infrequente volgorde van activiteiten redelijk lage scores toekent aan de grotere event logs. In vergelijking met de kleinere event logs, hebben deze grotere event logs ook meer traces en gemiddeld gezien minder cases per trace. Dit kan leiden tot een hoger aantal infrequente volgorde van activiteiten en deze score dus mogelijks verklaren. Verder valt op dat de metrieken die door meerdere tools gekwantificeerd zijn, in gelijke trend beoordeeld worden, met scores van Fischer et al. [11] die gemiddeld de laagste beoordeling geven. Ten slotte valt in de dimensie nauwkeurigheid nog op dat de 2017 event log activiteiten bevat met een negatieve duurtijd.

Bij de datakwaliteitsdimensie volledigheid valt op dat elke event log enorm veel ontbrekende events heeft. Ontbrekende events worden door Fischer et al. [11] als volgt gedetecteerd: In elke case wordt voor elke activiteit het aantal events geteld waar de status gelijk is aan 'begin' en het aantal events waar de status gelijk is aan 'compleet'. Indien deze niet gelijk zijn, leiden ze af dat de case een ontbrekende event voor die bepaalde activiteit bevat. Verder kan ook opgemerkt worden dat de metrieken ontbrekend tijdstip en event volgorde per trace door beide tools hetzelfde gekwantificeerd worden. Ten slotte valt

nog op dat het probleem, attribuut relevantie controle, voor de event log van 2017 redelijk slecht scoort. Deze metriek zoekt per attribuut op welke mogelijke invullingen slechts in minder dan 5% van de events voorkomt en ziet dit als een irrelevante attribuutwaarde. We kunnen dus afleiden dat de attributen in de 2017 event log heel veel verschillende mogelijke invullingen hebben. Indien een attribuutwaarde maar zo weinig voorkomt, heeft deze volgens Verhulst [9] geen belang voor proces mining op deze event log [9].

De consistentie datakwaliteitsdimensie wordt in het algemeen goed beoordeeld bij alle event logs. Enkel de granulariteit van de tijdstippen in de log, traces en activiteiten blijkt niet consistent gelogd te worden in de event log van 2018. Bij het meer gedetailleerd bekijken van de event log lijkt dit ook te kloppen, sommige tijdstippen zijn specifiek gelogd tot de milliseconde, terwijl andere tijdstippen slechts tot op de seconde of tot op de dag specifiek gelogd zijn. De consistentie controle scoort in het algemeen ook relatief laag ten opzichte van de andere metrieken. Deze metriek wordt beoordeeld door te kijken naar de lengte en de structuur van een attribuutwaarde. Er wordt gekeken of de lengte sterk verschilt van de andere mogelijke waarden en of de waarde is opgebouwd uit enkel tekens, nummers of een combinatie van beiden. Een perfecte score kan op deze metriek bereikt worden indien de attribuutwaarden consistent zijn qua lengte en telkens bestaan uit enkel tekens of nummers.

Ten slotte is hetgeen dat het meest opvalt bij de datakwaliteitsdimensie uniekheid dat de duplicaten in de event log metrieken door Fischer et al. [11] en Verhulst [9] niet compleet overeenkomen. Volgens de kwantificering van Verhulst [9] hebben de 2018 event log en de event log van 2020-DD geen duplicaten in de event log, terwijl dit volgens Fischer et al. [11] wel het geval is. Wel zijn beide methodes akkoord dat de event log van 2019 heel veel duplicaten bevat en geven ze scores in gelijkaardige trend.

Kort samengevat is een duidelijk verschil in de datakwaliteit op te merken tussen de verschillende event logs. De event log van 2017 scoort voornamelijk hoog op de datakwaliteitsdimensies consistentie en uniekheid. De 2017 event log scoort duidelijk hoger op de uniekheid dimensie ten opzichte van de andere event logs. De 2018 event log scoort over het algemeen het slechtst op alle dimensies, enkel voor de volledigheid dimensie scoort de 2020-RFP event log lager. Verder scoren beide 2020 event logs hoge scores op de andere datakwaliteitsdimensies. Echter blijft de gemiddelde datakwaliteit van de 2017 event log het hoogste. De 2019 event log scoort gemiddeld op alle datakwaliteitsdimensies en heeft voor geen van de beoordeelde dimensies de hoogste of laagste beoordeling.

5 Discussie

De empirische studie vergelijkt een set van vijf diverse publieke real-live event logs met elkaar op vlak van datakwaliteit aan de hand van vier verschillende datakwaliteitsdimensies. De event log kwaliteit is in

het algemeen hoog beoordeeld. Echter is het belangrijk op te merken dat in elk van de publieke real-life event logs diverse datakwaliteitsproblemen aanwezig zijn. Deze datakwaliteitsproblemen zijn daarenboven niet consistent doorheen alle event logs. De datakwaliteit van geen van de onderzochte event logs is perfect, dus publieke event logs zijn ook onderhevig aan datakwaliteitsproblemen. Deze problemen moeten geïdentificeerd en meegenomen worden bij het uitvoeren van proces mining en andere technieken die toegepast worden op deze publieke event logs. Deze benchmarking studie kan een indicatie geven van de datakwaliteitsproblemen in deze event logs en daarenboven kunnen nieuwe publieke real-life event logs beoordeeld worden met de toegepaste methode van de empirische studie.

Echter is het belangrijk op te merken dat er aan deze empirische studie ook enkele beperkingen verbonden zijn. Zo zijn in deze empirische studie niet alle datakwaliteitsdimensies beoordeeld. Het is echter niet mogelijk alle dimensies te beoordelen met de event log als enige input. Nadat toekomstig onderzoek het mogelijk maakt om meerdere dimensies te kwantificeren op basis van enkel de event log, kunnen deze dimensies toegevoegd worden aan de studie om een vollediger overzicht te geven van de datakwaliteitsproblemen in de event logs.

Ten tweede is in deze empirische studie slechts gebruik gemaakt van drie verschillende tools om de datakwaliteit van de event logs te beoordelen (Fischer et al. [11], Verhulst [9], en DaQAPO [21]). Dit zijn alle beschikbare tools om de besproken taxonomieën (in sectie 3.4) toe te passen op event logs. Het kan echter interessant zijn om meerdere tools op te nemen in de selectie wanneer deze ontwikkeld zijn (of zelf ontwikkeld worden). Het opnemen van meerdere tools maakt het mogelijk om de resultaten van de andere tools te ondersteunen en daarenboven zullen per gekwantificeerde dimensies meerdere metrieken gebruikt worden in de beoordeling. Dit kan leiden tot een meer complete beoordeling van de event logs en zo de kwaliteit van de beoordeling verder verhogen.

Ten derde is de benchmarking studie uitgevoerd op een set van vijf event logs. Ondanks dat er in deze set event logs veel variatie zat, zijn er nog veel meer verschillende soorten event logs om met elkaar te vergelijken. Het uitvoeren van deze studie op een grotere set event logs zou de studie completer maken alsook een vollediger overzicht bieden om nieuwe event logs mee te vergelijken. Uiteindelijk had geen enkele beoordeelde event log een extreem lage kwaliteit. Toch kan het zeker interessant zijn om ook een event log met zeer lage datakwaliteit toe te voegen aan deze selectie.

Ten slotte is de datakwaliteit van deze event logs louter en alleen bepaald op basis van de event log zelf, dus zonder invloed van domeinexperten of andere inputs. Dit was ook de focus van deze masterproef, maar in een industriële context gaat het vaker voorkomen dat andere inputs wel ter beschikking zijn, dewelke dan ook andere interessante inzichten kunnen leveren. Verder onderzoek kan mogelijks focussen op het beoordelen van de datakwaliteit van event logs in contexten waar wel externe inputs beschikbaar zijn.

6 Conclusie

De datakwaliteit van de gebruikte input event log is belangrijk om correcte en betrouwbare proces mining resultaten te bekomen. Om een methode aan te tonen om publieke event logs te beoordelen en een overzicht te verkrijgen van de datakwaliteit van deze event logs, is een empirische studie uitgevoerd op een set van vijf diverse event logs. Om in de empirische studie een betrouwbare analyse uit te kunnen voeren, is eerst onderzoek gedaan naar de bestaande taxonomieën, datakwaliteitsdimensies en imperfectie patronen die datakwaliteitsproblemen identificeren en beoordelen. Op basis van de verworven inzichten is een set van datakwaliteitsdimensies samengesteld om de datakwaliteit van vijf event logs te beoordelen. Deze dimensies zijn nauwkeurigheid, volledigheid, consistentie, en uniekheid. Het was belangrijk dat alle geselecteerde datakwaliteitsdimensies te beoordelen waren enkel op basis van de event log. Elk van de vier dimensies is gekwantificeerd met behulp van een set van metrieken. Deze metrieken zijn toegepast op een set van vijf diverse publieke real-life event logs van de BPI challenge. De datakwaliteit van deze event logs is bepaald op basis van het R-pakket DaQAPO en twee geïmplementeerde extensies in ProM, Fischer et al. [11] en Verhulst [9].

Na het beoordelen van de datakwaliteit van de event logs viel op dat alle beoordeelde event logs onderhevig waren aan diverse datakwaliteitsproblemen. Deze datakwaliteitsproblemen zijn daarenboven niet consistent tussen de beoordeelde event logs. Hieruit valt af te leiden dat de beschikbare publieke real-life event logs beoordeeld moeten worden om datakwaliteitsproblemen te identificeren zodat deze opgeschoond of meegenomen kunnen worden bij het uitvoeren van de process mining analyses.

Toekomstig onderzoek kan focussen op het kwantificeren van datakwaliteitsdimensies die momenteel nog niet te beoordelen zijn zonder externe input en deze toe te voegen aan de empirische studie. Verder is het interessant om een bredere set tools toe te passen op de event logs om zo de andere tools te ondersteunen en per datakwaliteitsdimensie meerdere metrieken te beoordelen. Ten derde kunnen meerdere event logs toegevoegd worden aan de set van beoordeelde event logs om zo een completer beeld te geven over de verschillende event logs. Ten slotte kan toekomstig onderzoek ook focussen op het beoordelen van de datakwaliteit van event logs waar meerdere inputs beschikbaar zijn.

7 Bijlagen

7.1 Maturiteitsniveaus van het Process Mining Manifesto [30]

Niveau	Uitleg
Eén	Een event log waarin events doorgaans niet automatisch worden vastgelegd. Opgenomen events komen mogelijk niet overeen met de werkelijkheid en events kunnen ontbreken.
Twee	Een event log waarin events automatisch worden vastgelegd door een bepaald informatie-systeem, maar waar een systematische aanpak voor logging ontbreekt. Bovendien kan het systeem worden omzeild. Als gevolg hiervan kunnen events ontbreken of niet overeenkomen met de werkelijkheid.
Drie	Een event log waarin events automatisch worden vastgelegd, maar geen systematische aanpak wordt gevolgd voor logging. Hoewel er misschien events ontbreken, is er een redelijk vertrouwen dat de geregistreerde events overeenkomen met de werkelijkheid, in tegenstellingen tot event logs van maturiteitsniveau twee.
Vier	Een event log waarin events automatisch en systematisch worden vastgelegd, met andere woorden de inhoud van de event log is zowel betrouwbaar als volledig. Bovendien zijn er expliciete casus- en activiteit- begrippen aanwezig.
Vijf	Een event log van uitstekende kwaliteit, betrouwbaar en volledig, en events zijn goed gedefinieerd. Events worden op een automatische, systematische, betrouwbare en veilige manier vastgelegd. Bovendien zijn privacy- en beveiligingsproblemen naar behoren behandeld.

Tabel 7: Maturiteitsniveaus van het Process Mining Manifesto [30]

7.2 Datakwaliteitsproblemen van Bose et al. [6]

Ontbrekende data	
Ontbrekende cases	Cases die niet voorkomen in de event log, maar wel gebeurd zijn. Hierdoor is het mogelijk dat de process mining resultaten niet de realiteit weerspiegelen.
Ontbrekende events	Events in een trace die niet opgenomen zijn. Hierdoor kunnen relaties gevonden worden die eventueel niet in de realiteit voorgekomen zijn.
Ontbrekende relaties	De afwezigheid van een expliciete link tussen een event en een case. Dit maakt het moeilijker om de juiste relaties tussen verschillende events af te leiden.
Ontbrekende case attributen	Case attributen die niet gelogd zijn. Algoritmes moeten deze specifieke case laten vallen indien deze waarde gebruikt wordt.
Ontbrekende positionering	Het is niet geweten op welk moment het event voorkomt in de trace, enkel van toepassing indien ook het tijdstip ontbreekt.
Ontbrekende activiteit labels	De oorsprong van een event is onduidelijk omdat de activiteit naam niet gelogd is.
Ontbrekende tijdstippen	Er is geen tijdstip gelogd, wat performance analyses meer complex maakt. Bovendien is ontdekte control-flow niet betrouwbaar indien de positionering mogelijks incorrect is.
Ontbrekende resources	De resource van een event is niet geregistreerd waardoor process mining algoritmes die hiervan gebruik maken meer moeite ondervinden.
Ontbrekende event attributen	Een event attribuut is niet geregistreerd waardoor process mining algoritmes die gebruik maken van de deze attributen meer moeite ondervinden.
Incorrecte data	

Vervolg Tabel 8 van vorige pagina

Incorrecte cases	Cases in een event logs van een proces die in de realiteit behoren tot een ander proces. Dit zorgt voor uitschieters in de process mining analyses.
Incorrecte events	Events die niet zijn gebeurd maar wel voorkomen in de event log.
Incorrecte relaties	Een foute link tussen een event en een case waardoor deze events aan de foute cases gelinkt worden.
Incorrecte case attributen	Case attributen die fout gelogd zijn. Dit maakt het moeilijker voor process mining algoritmes die deze attribuutwaarden gebruiken.
Incorrecte positionering	Een event is fout gepositioneerd in de trace, enkel van toepassing indien ook het tijdstip ontbreekt.
Incorrecte activiteit labels	De naam van de activiteit is fout gelogd.
Incorrecte tijdstippen	Het tijdstip komt niet overeen met het exacte tijdstip waarop het event zich in de realiteit heeft afgespeeld.
Incorrecte resources	Een resource die foutief gelinkt is aan een event waardoor algoritmes die deze informatie gebruiken problemen ondervinden.
Incorrecte event attributen	Een event attribuut dat fout geregistreerd is waardoor process mining algoritmes die deze informatie gebruiken problemen ondervinden.
Onnauwkeurige data	
Onnauwkeurige relaties	Foute keuze over de gebruikte definitie van een case. Hierdoor kan extra, eventueel belangrijke, informatie verloren gaan.
Onnauwkeurige case attributen	De invulling van de case attributen waarden is niet specifiek genoeg waardoor extra informatie verloren gaat.
Onnauwkeurige positionering	De positionering van bepaalde events is fout. Bijvoorbeeld in het geval dat twee activiteiten in de realiteit in parallel lopen, maar door de positionering als sequentieel gelogd worden. Hierdoor kan de correcte control-flow moeilijker of niet bepaaldworden door algoritmen. Dit is enkel van toepassing indien ook de tijdstippen ontbreken.
Onnauwkeurige activiteit labels	De activiteit namen zijn niet specifiek genoeg waardoor meerdere verschillende activiteiten onder eenzelfde naam vallen.
Onnauwkeurige tijdstippen	Alle of een gedeelte van de tijdstippen zijn niet specifiek genoeg gelogd waardoor het niet duidelijk is welk event eerder is voorgekomen in de realiteit.
Onnauwkeurige resources	Meer specifieke informatie over een resource is geweten maar niet gelogd in de event log. Hierdoor zijn inzichten over resources gelimiteerd bij het toepassen van process mining.
Onnauwkeurige event attributen	De waarden voor event attributen zijn niet specifiek genoeg gelogd waardoor extra informatie verloren gaat. Bijvoorbeeld: bij de temperatuur attribuut worden enkel de volgende waarden geaccepteerd: 'vries' (<= 0°C), 'koud' (0°C - 18°C), en 'warm' (>= 18°C). Hierdoor zal het eventueel moeilijker zijn om het exacte effect van de temperatuur in te schatten.
Irrelevante data	
Irrelevante cases	De in de event log opgenomen cases die niet relevant zijn voor de analyse die wordt uitgevoerd. Dit kan de begrijpelijkheid van de process mining resultaten negatief beïnvloeden.
Irrelevante events	De events die niet relevant zijn voor de uit te voeren analyse. De events zullen nog gefilterd en geaggregeerd moeten worden.

Tabel 8: Datakwaliteitsproblemen van Bose et al. [6]

7.3 Datakwaliteitsdimensies van Verhulst [9]

Dimensie	Uitleg
Volledigheid	De data bevatten alle nodige informatie, er zijn geen ontbrekende waarden en de transactionele data is aanwezig.
Uniekheid of Duplicaten	Er zijn geen traces met exact dezelfde attributen meerdere keren aanwezig in de event log.
Tijdigheid	De event log bevat enkel events in de verwachte tijdsperiode.
Geldigheid	De data voldoen aan de syntactische (formaat, type, range) vereisten van zijn definities.
Nauwkeurigheid of correctheid	De gelogde waarden voldoende overeen met de waarden in de realiteit.
Consistentie	De waarden in de gehele event log zijn consistent opgenomen.
Geloofwaardigheid	De data worden als betrouwbaar en objectief gezien door de gebruikers.
Relevantie	De data zijn relevant voor de vooropgestelde analyses.
Veiligheid of vertrouwelijkheid	De data kunnen veilig worden behandeld.
Complexiteit	Deze dimensie beschrijft de complexiteit van de data (het aantal subtraces aanwezig, ...)
Samenhang	Er is een logische interconnectie tussen data inputs.
Representatie of formaat	De data zijn compact en in eenzelfde formaat gepresenteerd.

Tabel 9: Datakwaliteitsdimensies van Verhulst [9]

7.4 Imperfectie patronen van Suriadi et al. [27]

Dimensie	Uitleg
Op formulieren gebaseerde event registratie	Event data wordt verzamelt aan de hand van elektronische formulieren in een informatiesysteem. Werken met formulieren zorgt ervoor dat meerdere events worden vastgelegd met eenzelfde tijdstip wanneer het formulier in het systeem wordt ingediend. Echter bestaat de kans dat de onderliggende events op verschillende tijdstippen zijn uitgevoerd.
Onbedoeld tijdreizen	Met onbedoeld tijdreizen worden events bedoelt die een tijdstip krijgen dat gelijkend is op het echte tijdstip. In het geval dat een activiteit enkele minuten na middernacht plaatsvindt, is het mogelijk dat de werknemer het juiste uur ingeeft, maar de datum van de dag er voren.
Inconsistent tijdsformaat	Tijdstippen worden vastgelegd in een ander formaat dan het door de tooling verwachtte formaat.
Verspreid event	Indien informatie in de attribuutwaarden de aanwezigheid van aanvullende events benadrukken. Deze events worden echter niet expliciet vastgelegd in aparte events, maar zijn verborgen in de attribuutwaarden van een ander event die in de event log worden vastgelegd.
Niet gelinkt event	Dit probleem doet zich voor wanneer events niet aan een case zijn gekoppeld, zoals vaak wordt waargenomen wanneer data afkomstig zijn van informatiesystemen die niet proces gefocust zijn.

Vervolg Tabel 10 van vorige pagina

Dimensie	Uitleg
Verspreide case	Een case waarvoor kernactiviteiten niet zijn vastgelegd in eenzelfde event log. Om inzicht te krijgen in de volledige processtroom voor deze case, moet de event log worden samengesteld met behulp van de inhoud van meerdere informatiesystemen.
Collateral event	*Collateral events* zijn verschillende events die verwijzen naar dezelfde proces activiteit. *Collaterale events* worden veroorzaakt op drie manieren: (1) het event log is opgesteld op basis van meerdere systemen, die elk op hun eigen manier eenzelfde proces activiteit vastleggen, (2) het gebruikte onderliggende systeem voert automatisch een reeks hulp events uit wanneer een specifiek event plaatsvindt, en/of (3) de event log legt gebruikersactiviteiten gedetailleerd op een laag niveau vast waardoor dubbele events binnen een zeer korte tijdsperiode geregistreerd worden (bijvoorbeeld indien het openen en sluiten van een formulier wordt geregistreerd en de gebruiker enkele keren heen en weer wisselt tussen twee formulieren).
Vervuild label	Verschillende event attribuutwaarden hebben dezelfde structuur, maar verschillen in termen van hun specifieke waarden. Wanneer een dergelijk event log zou worden gebruikt voor detectie van control-flows, zou het lijden onder het grote aantal unieke activiteiten labels.
Vervormd label	Vervormde labels zijn waarden van meerdere event attributen die niet identiek zijn, maar die zeer sterke syntactische en semantische overeenkomsten vertonen.
Label synoniemen	Dit probleem treedt op wanneer verschillende waarden op syntactisch niveau verschillen, maar op semantisch niveau vergelijkbaar zijn. Bijvoorbeeld: in twee verschillende informatiesystemen wordt naar dezelfde activiteit verwezen met twee verschillende labels.
Gelijknamige labels	Een activiteit die wordt herhaald voor een bepaalde case, maar de semantiek van deze activiteit is niet hetzelfde in beide situaties.

Tabel 10: Imperfectie patronen van Suriadi et al. [27]

7.5 Datakwaliteitsproblemen van Kherbouche et al. [15]

Complexiteit	
Structurele complexiteit	Verwijst naar de syntactische middelen (bijvoorbeeld parallelle activiteiten, loops, ...) die nodig zijn om een procesmodel op te stellen van de event log. De structurele complexiteit wordt gekwantificeerd op basis van het bestaan van loops, dubbele taken, verborgen taken, en de hoeveelheid events en traces in de event log.
Gedrags-complexiteit	Verwijst naar de complexiteit van het aanwezige gedrag in de event log. De gedragscomplexiteit wordt bepaalt aan de hand van hoe erg de verschillende traces in een event log van elkaar verschillen. Ook de hoeveelheid en de complexiteit van de events in traces worden hierbij bekeken
Nauwkeurigheid	
Precisie	Meet het niveau van nauwkeurigheid of precisie van event data aanwezig zijn in de event log. De precisie van event data wordt bepaalt aan de hand van de events, traces, tijdstippen en/of resources.

Vervolg Tabel 11 van vorige pagina

Betrouwbaarheid	Zegt dat de geregistreerde events ook daadwerkelijk moeten hebben plaatsgevonden en dat de attributen van de events correct moeten zijn. De betrouwbaarheid wordt bepaalt op basis van de oorsprong(en) van de event data, betrouwbaarheid van de generatie van de data (automatisch of handmatig), en de betrouwbaarheid van de oorspronkelijke event data bron(nen).
Consistentie	
Correctheid	Evalueert hoe correct de data in de event log geschat wordt (bijvoorbeeld doordat er geen uitschieters in de data aanwezig is)
Integriteit	Meet de geldigheid van event data die kunnen worden aangetast door menselijke fouten wanneer data worden ingevoerd, geïmporteerd worden van heterogene event log bronnen of hardware storingen. Elk event moet verwijzen naar een case en elke case bestaat uit een opeenvolging van juist geordende events.
Gestructureerdheid	Verwijst naar de vorm van de event data. De event data moeten beschikbaar zijn in een gestructureerd formaat waarmee direct gewerkt kan worden en niet bijvoorbeeld in vrije tekst vorm.
Volledigheid	
Beschikbaarheid	Verwijst naar de beschikbaarheid van de vereiste event data in de event log. Ontbrekende data leveren mogelijks een resultaat op dat niet overeenkomt met de werkelijke uitvoering van de cases of maken het onmogelijk om bepaalde, specifieke soorten analyses toe te passen (bijvoorbeeld organizational miner, social network miner).
Lokale volledigheid	Meet of alle directe opvolging relaties tussen taken in de event log zijn vastgelegd.
Globale volledigheid	Meet of al het mogelijke gedrag van het procesmodel opgenomen is in de event log.

Tabel 11: Datakwaliteitsproblemen van Kherbouche et al. [15]

7.6 Datakwaliteitsproblemen van Vanbrabant et al. [33]

Ontbrekende data	
Ontbrekende waarden	Ontbrekende waarden zijn data waarden die aanwezig zouden moeten zijn, maar die niet worden vastgelegd. Daarbij moet onderscheid worden gemaakt tussen werkelijk ontbrekende waarden (die wel gelogd hadden moeten worden) en waarden waarvan het logisch en verwacht is dat er geen waarde is gelogd.
Ontbrekende attributen	Attributen die nodig zijn voor de analyse, maar die niet opgenomen zijn in de event log. Terwijl ontbrekende waarden specifieke data weerspiegelen die voor bepaalde patiënten ontbreken, impliceren ontbrekende attributen dat een attribuutwaarde voor alle patiënten ontbreekt.
Ontbrekende cases	Ontbrekende cases houdt in dat cases die in de realiteit zijn afgehandeld, niet in de event log worden weergegeven.
Geschonden attribuut afhankelijkheden	
Schending van de logische volgorde	Een schending van de logische volgorde houdt in dat de volgorde van bepaalde activiteiten onjuist is vanwege problemen met de geregistreerde tijdstippen.

Vervolg Tabel 12 van vorige pagina

Schending van wederzijdse afhankelijkheid	Een schending van wederzijdse afhankelijkheid treedt op als twee van elkaar afhankelijke attributen tegenstrijdige waarden hebben.
Incorrecte attribuut waarden	
Onnauwkeurigheid van tijdstippen	Onnauwkeurige tijdstippen zijn tijdstippen die niet exact de correcte tijd weergeven waarop een event in realiteit plaatsvond. Dit is een veelvoorkomend probleem met de datakwaliteit indien tijdstippen worden vastgelegd na een handmatige actie door een werknemer. Wanneer er een discrepantie is tussen het moment waarop een actie wordt uitgevoerd en het moment waarop deze in het systeem wordt vastgelegd, treedt dit datakwaliteitsprobleem op.
Typefouten	Typefouten in tekstvelden kunnen ook fouten in attribuutwaarden veroorzaken, wat leidt tot problemen voor algoritmen die deze attributen gebruiken.
Buiten bereik van het domein	Schendingen van domein bereik verwijzen naar tijdstippen, numerieke en categorische waarden die buiten het bereik van mogelijke waarden vallen.
Andere onwaarschijnlijke waarden	Dit probleem is een restcategorie van verkeerde waarden die niet overeenkomen met een van de eerdere specificaties.
Niet foute, maar niet direct bruikbare data	
Inconsistent formaat	De waarden van een of meerdere attributen van hetzelfde type is niet consistent binnen een case of tussen meerdere cases.
Impliciete waarde nodig	Dit kwaliteitsprobleem verwijst naar attribuutwaarden die niet expliciet beschikbaar zijn, maar die kunnen worden berekend of afgeleid uit andere, wel beschikbare data.
Ingesloten waarden	Ingesloten waarden zijn attribuutwaarden die een samentrekking zijn van meerdere andere delen bruikbare informatie. Wanneer het formaat van deze samentrekking consistent is, kunnen de ingesloten waarden worden verkregen door deze samentrekking te splitsen.
Afkortingen	Afkortingen worden vaak gebruikt om specifieke domein terminologie in te korten. Vanuit een analyse perspectief kunnen afkortingen problematisch zijn, vooral wanneer ze niet consequent worden gebruikt.
Onnauwkeurige data	Onnauwkeurige data verwijzen naar data die niet gedetailleerd genoeg zijn opgenomen.

Tabel 12: Datakwaliteitsproblemen van Vanbrabant et al. [33]

7.7 Datakwaliteitproblemen van Fischer et al. [11]

Nauwkeurigheid	
Infrequente volgorde van activiteiten	Deze metriek detecteert onregelmatige volgordes van activiteiten. Indien twee activiteiten in de meeste gevallen elkaar in eenzelfde volgorde opvolgen, zullen de enkele minder frequente relaties als infrequente volgorde beschouwd worden. Ook onbedoeld tijdreizen zal hierdoor worden opgevangen omdat dit zich normaal voordoet in zeldzame situaties waarin de activiteiten ordening duidelijk afwijkend is.

Vervolg Tabel 13 van vorige pagina

Overlappende activiteiten per resource	Indien de start- en eindtijden van een activiteit die wordt uitgevoerd door een resource worden vastgelegd in een event log, is het mogelijk om te detecteren wanneer een resource op eenzelfde moment meerdere activiteiten uitvoert. Gegeven de assumptie dat een resource niet multitaskt kan dit duiden op onnauwkeurige registratie van start- en/of eindtijd van activiteiten aangezien de metriek activiteiten identificeert die door een specifieke resource zijn gestart voordat die hun vorige activiteit beëindigd hadden.
Toekomstige invoer	Events met tijdstippen in de toekomst worden gedetecteerd met deze metriek. Omdat event logs een overzicht van verleden events vertegenwoordigt, worden toekomstige tijdstippen als een probleem in event logs beschouwd.
Granulariteit	Met deze metriek worden events met te onnauwkeurige tijdstippen opgespoord. Handmatig gelogde events vertonen vaak een te onnauwkeurige granulariteit, omdat het voor de logger moeilijk is om dit tot op de milliseconde precies vast te leggen.
Volledigheid	
Ontbrekende trace	Indien het verschil tussen de tijdstippen van de start events van twee opeenvolgende traces significant groter is dan de verwachte gemiddelde waarde wordt dit gezien als indicator voor een ontbrekende trace. Daarenboven wordt ook het tijdsverschil tussen het begin van de dag en het start event van de eerste trace bekeken. Dit gebeurt eveneens voor het start event van de laatste trace van de dag en het einde van de dag.
Ontbrekende activiteit	De metriek beschrijft de situatie waarin een activiteit verwacht wordt in een trace maar niet aanwezig is. Enkele van de mogelijke oorzaken hiervoor zijn dat men vergeten is de activiteit te loggen of dat bepaalde delen van het systeem niet verbonden zijn.
Ontbrekend event	Cases waarbij een activiteit ofwel een start- of eind event mist worden gedetecteerd door deze metriek. Een case is enkel gezien als consistent als elk start event een overeenkomstig eind event heeft en omgekeerd. Een mogelijke reden is het weglaten van het start of eind event of indien het verwachte event toegewezen aan een verkeerde case.
Ontbrekend tijdstip	Deze metriek detecteert events zonder tijdstip. Ook tijdstippen voor het jaar 1971 worden als ontbrekend tijdstip gezien omdat de meeste systemen tijd waardes omzetten naar UNIX tijdstippen.
Conistentie	
Formaat	Een event log kan zowel tijdstippen bevatten die opgeslagen zijn in een dag-maand-jaar indeling als tijdstippen die opgeslagen zijn in een maand-dag-jaar indeling indien de event log bijvoorbeeld opgebouwd is vanuit meerdere systemen. Hierdoor zullen de tijdstippen van events niet consistent vastgelegd zijn, wat opgemerkt wordt door deze metriek.
Gemengde granulariteit van de event log	Deze metriek controleert of bepaalde events meer of minder nauwkeurig worden vastgelegd dan de rest. Dit kan onder meer voorkomen als bepaalde delen van de event log automatisch worden vastgelegd via elektronische systemen terwijl andere delen handmatig door de gebruiker worden geregistreerd.
Gemengde granulariteit van de traces	Er wordt gecontroleerd of de tijdstippen van bepaalde cases meer of minder nauwkeurig worden geregistreerd. Dit wordt mogelijks veroorzaakt wanneer enkel bepaalde cases handmatig geregistreerd worden (omdat elektronische systeem registratie eventueel niet mogelijk is).

Vervolg Tabel 13 van vorige pagina

Gemengde granulariteit van de activiteiten	Er wordt gecontroleerd of bepaalde activiteiten in de event log meer of minder nauwkeurig worden geregistreerd. Dit kan voorkomen indien slechts bepaalde activiteiten handmatig geregistreerd worden (omdat elektronische systeem registratie eventueel niet mogelijk is).
Uniekheid	
Duplicaten in de event log	Deze metriek detecteert events met exact hetzelfde tijdstip, maar die toch behoren tot verschillende cases. Het loggen van events via e-formulieren kan ervoor zorgen dat meerdere events opgeslagen worden op eenzelfde tijdstip, terwijl deze niet gelijktijdig plaatsvonden.
Duplicaten in een trace	Deze metriek detecteert events met exact hetzelfde tijdstip binnen eenzelfde case. Net zoals bij duplicaten in de event log kan dit veroorzaakt worden wanneer events via e-formulieren gelogd worden. Deze worden dan eventueel opgeslagen met identieke tijdstippen ondanks deze niet gelijktijdig plaatsvonden.
Duplicaten in een activiteit	Events met exact hetzelfde tijdstip en binnen dezelfde activiteit worden met deze metriek gekwantificeerd. Start en eind events met eenzelfde tijdstip zijn een typisch voorbeeld voor dit probleem.

Tabel 13: Datakwaliteitsproblemen van Fischer et al. [11]

7.8 Overzicht datakwaliteitsproblemen en imperfectie patronen in de dimensies van Fischer et al. [11]

Dimensie	Kwantificeerbaar	Datakwaliteitsproblemen en imperfectiepatronen
Beschikbaarheid	Nee	/
Nauwkeurigheid	Ja	**Bose et al. [6]** Incorrecte case attributen Incorrecte positionering Incorrecte activiteit labels Incorrecte tijdstippen Incorrecte resources Incorrecte event attributen Onnauwkeurige case attributen Onnauwkeurige positionering Onnauwkeurige activiteit labels Onnauwkeurige tijdstippen Onnauwkeurige resources Onnauwkeurige event attributen
		Kherbouche et al. [15] Precisie Correctheid
		Suriadi et al. [27] Onbedoeld tijdreizen
		Vanbrabant et al. [33] Schending van de logische volgorde Onnauwkeurigheid van tijdstippen Typefouten Onnauwkeurige data

Vervolg Tabel 14 van vorige pagina

Dimensie	Kwantificeerbaar	Datakwaliteitsproblemen en imperfectiepatronen
Nauwkeurigheid	Ja	**Fischer et al. [11]** Infrequente volgorde van activiteiten Overlappende activiteiten per resource Toekomstige invoer Granulariteit
Hoeveelheid data	Nee	/
Geloofwaardigheid	Nee	**Vanbrabant et al. [33]** Schending van wederzijdse afhankelijkheid Buiten bereik van het domein Andere onwaarschijnlijke waarden
Volledigheid	Ja	**Bose et al. [6]** Ontbrekende cases Ontbrekende events Ontbrekende case attributen Ontbrekende positionering Ontbrekende activiteit labels Ontbrekende tijdstippen Ontbrekende resoruces Ontbrekende event attributen
		Vanbrabant et al. [33] Ontbrekende waarden Ontbrekende attributen Ontbrekende cases
		Fischer et al. [11] Ontbrekende trace Ontbrekende activiteit Ontbrekend event Ontbrekend tijdstip
Beknoptheid	Nee	/
Consistentie	Ja	**Kherbouche et al. [15]** Integriteit Gestructureerdheid
		Suriadi et al. [27] Inconsistent formaat *Collateral events* Vervuild label Vervormd label Label synoniemen
		Vanbrabant et al. [33] Inconsistent formaat Afkortingen
		Fischer et al. [11] Gemengde granulariteit van de event log Formaat Gemengde granulariteit van de traces Gemengde granulariteit van de activiteiten
Data management	Nee	**Suriadi et al. [27]** Verspreide case

Vervolg Tabel 14 van vorige pagina

Dimensie	Kwantificeerbaar	Datakwaliteitsproblemen en imperfectiepatronen
Interlinking	Nee	**Bose et al. [6]** Ontbrekende relaties Incorrecte relaties
		Kherbouche et al. [15] Lokale volledigheid
		Suriadi et al. [27] Eenzaam event
Interpreteerbaarheid	Nee	/
Objectiviteit	Nee	/
Presentatie kwaliteit	Nee	/
Prijs	Nee	/
Relevantie	Nee	**Bose et al. [6]** Incorrecte cases Irrelevante cases Irrelevante events
		Suriadi et al. [27] Verspreid event
		Vanbrabant et al. [33] Impliciete waarde nodig Ingesloten waarden
Reputatie	Nee	/
Beveiliging	Nee	/
Service kwaliteit	Nee	/
Specialisatie	Nee	/
Tijdigheid	Nee	/
Traceerbaarheid	Nee	**Bose et al. [6]** Incorrecte events
		Suriadi et al. [27] Op formulieren gebaseerde event registratie
Begrijpelijkheid	Nee	**Kherbouche et al. [15]** Structurele complexiteit Gedragscomplexiteit
		Suriadi et al. [27] Gelijknamige labels
Uniekeid	Ja	**Fischer et al. [11]** Duplicaten in de event log Duplicaten in een trace Duplicaten in een activiteit
Bruikbaarheid	Nee	/
Waarde-toevoegend	Nee	**Bose et al. [6]** Onnauwkeurig relaties
Verifieerbaarheid	Nee	/

Tabel 14: Datakwaliteitsproblemen en imperfectie patronen in de dimensies van Fischer et al. [11]

7.9 Overzicht van de toegepaste datakwalieitsdimensies en metrieken

Dimensies en metrieken per auteur			**Uitleg van dimensies en metrieken**
Fischer et al. [FI]	Verhulst [VE]	DaQAPO [Da]	
Nauwkeurigheid			De waarden in de event log zijn representatief voor de waarden in realiteit.
Infrequente volgorde van activiteiten			Bij activiteiten die in de meeste gevallen elkaar in eenzelfde volgorde opvolgen, zullen de minder frequente relaties als fout beschouwd worden.
Overlappende activiteiten per resource		Overlappende activiteit	Een resource die op eenzelfde moment meerdere activiteiten uitvoert.
Toekomstige invoer	Tijdstip correctheid		Events met tijdstippen in de toekomst.
Granulariteit	Tijdstip formaat		Events met onnauwkeurige tijdstippen.
		Negatieve duur van activiteiten	Events met een compleet tijdstip dat zich afspeelt voor het start tijdstip.
Volledigheid			Alle waardes voor één specifieke variabele zijn opgenomen in de event log.
Ontbrekende trace			Het verschil tussen de tijdstippen van de start events van twee opeenvolgende traces is significant groter is dan de verwachte gemiddelde waarde.
Ontbrekende activiteit			Een activiteit wordt verwacht in een trace maar is niet aanwezig.
Ontbrekend event			Een case waarbij een activiteit ofwel een start- of eind event ontbreekt.
Ontbrekend tijdstip	Event volgorde per trace		Events zonder tijdstip.
	Ontbrekende waarden		Missende waarden in alle attributen.
	Event-Resource controle		Events hebben geen resource.
	Transactionele informatie controle		Status informatie voor events is niet gelogd (in beste scenario is status informatie aanwezig met twee status per event).
	Attribuut relevantie controle		Zeldzame attribuutwaarden per attribuut.
Consistentie			De data waarden worden in alle events op een gelijkwaardige manier gelogd.
Formaat			De tijdstippen van events zijn niet consistent vastgelegd.
Gemengde granulariteit van de log			Tijdstippen van bepaalde events worden meer of minder nauwkeurig vastgelegd dan de rest.
Gemengde granulariteit van traces			Tijdstippen van bepaalde cases worden meer of minder nauwkeurig vastgelegd dan de rest.

Vervolg Tabel 15 van vorige pagina

Dimensies en metrieken per auteur			**Uitleg van dimensies en metrieken**
Fischer et al. [FI]	Verhulst [VE]	DaQAPO [Da]	
Consistentie			De data waarden worden in alle events op een gelijkwaardige manier gelogd.
Gemengde granulariteit van activiteiten			Tijdstippen van bepaalde activiteiten worden meer of minder nauwkeurig vastgelegd dan de rest.
	XES standaarden		De event log voldoet niet aan alle XES standaarden.
	Consistentie controle		Op basis van de lengte van de strings en of de string telkens bestaat uit enkel tekens of nummers of een combinatie, voor een attribuut vinden meerdere combinaties plaats.
Uniekheid			Geen ongewenste duplicaten aanwezig in de event log.
Duplicaten in de event log	Duplicaten in de event log		Events met exact hetzelfde tijdstip, maar die toch behoren tot verschillende cases.
Duplicaten in een trace	Duplicaten in een trace		Events met exact hetzelfde tijdstip binnen eenzelfde case.
Duplicaten in een activiteit			Events met exact hetzelfde tijdstip en binnen dezelfde activiteit.

Tabel 15: Overzicht van de toegepaste datakwaliteitsdimensies en metrieken

8 Referenties

[1] Andrews, R., Suriadi, S., Ouyang, C. & Poppe, E. (2018). Towards Event Log Querying for Data Quality. *Lecture Notes in Computer Science*.

[2] Andrews, R., Wynn, M., Vallmuur, K., ter Hofstede, A., Bosley, E., Elcock, M. & Rashford, S. (2019). Leveraging Data Quality to Better Prepare for Process Mining: An Approach Illustrated Through Analysing Road Trauma Pre-Hospital Retrieval and Transport Processes in Queensland. *International Journal of Environmental Research and Public Health*.

[3] Augusto, A., Conforti, R., Dumas, M., Rosa, M., Maggi, F., Marrella, A., Mecella, M. & Soo, A. (2019). Automated Discovery of Process Models from Event Logs: Review and Benchmark. *IEEE Transactions on Knowledge and Data Engineering*.

[4] Batini, C., Cappiello, C., Francalanci, C. & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*.

[5] Bezerra, F. & Wainer, J. (2013). Algorithms for anomaly detection of traces in logs of process aware information systems. *Information Systems*.

[6] Bose, R., Mans, R. & Van der Aalst, W. (2013). Wanna improve process mining results? *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*.

[7] Chinosi, M. & Trombetta, A. (2012). BPMN: An introduction to the standard. *Computer Standards & Interfaces*.

[8] Dixit, P., Suriadi, S., Andrews, R., Wynn, M., ter Hofstede, A., Buijs, J. & Van der Aalst, W. (2018). Detection and Interactive Repair of Event Ordering Imperfection in Process Logs. *Advanced Information Systems Engineering*.

[9] Eindhoven University of Technology & Verhulst. (2016). *Evaluating quality of event data within event logs an extensible framework*. https://research.tue.nl/en/studentTheses/evaluating-quality-of-event-data-within-event-logs

[10] Fernandez-Llatas, C. (2021). *Interactive Process Mining in Healthcare (Health Informatics)*. Springer.

[11] Fischer, D., Goel, K., Andrews, R., van Dun, C., Wynn, M. & Röglinger, M. (2022). Towards interactive event log forensics: Detecting and quantifying timestamp imperfections. *Information Systems*.

[12] Gschwandtner, T., Gärtner, J., Aigner, W. & Miksch, S. (2012). A Taxonomy of Dirty Time-Oriented Data. *Lecture Notes in Computer Science*.

[13] Janssenswillen, G., Depaire, B., Swennen, M., Jans, M. & Vanhoof, K. (2019). bupaR: Enabling reproducible business process analysis. *Knowledge-Based Systems*.

[14] Kahn, M., Raebel, M., Glanz, J., Riedlinger, K. & Steiner, J. (2012). A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Medical Care*.

[15] Kherbouche, M., Laga, N. & Masse, P. (2016). Towards a better assessment of event logs quality. *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*.

[16] Kim, W., Choi, B., Hong, E., Kim, S. & Lee, D. (2003). A taxonomy of Dirty Data. *Data Mining and Knowledge Discovery*.

[17] Kurniati, A., Rojas, E., Hogg, D., Hall, G. & Johnson, O. (2018). The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. *Health Informatics Journal*.

[18] Lu & Fahland. (2017). A conceptual framework for understanding event data quality in behavior analysis.

[19] Mans, R., Van der Aalst, W. & Vanwersch, R. (2015). Process Mining in Healthcare. *SpringerBriefs in Business Process Management*.

[20] Marin-Castro, H. & Tello-Leal, E. (2021). Event Log Preprocessing for Process Mining: A Review. *Applied Sciences*.

[21] Martin, N., Van Houdt, G. & Janssenswillen, G. (2022). DaQAPO: Supporting flexible and fine-grained event log quality assessment. *Expert Systems with Applications*.

[22] Müller & Freytag. (2003). Problems , Methods , and Challenges in Comprehensive Data Cleansing. *Humboldt University Berlin*.

[23] Nguyen, H., Lee, S., Kim, J., Ko, J. & Comuzzi, M. (2019). Autoencoders for improving quality of process event logs. *Expert Systems with Applications*.

[24] Oliveira, Rodrigues & Henriques. (2005). A Formal Definition of Data Quality Problems. *Proceedings of the 2005 International Conference on Information Quality*.

[25] Rahm & Do. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*.

[26] Ridley, D. (2012). *The Literature Review*. SAGE Publications.

[27] Suriadi, S., Andrews, R., ter Hofstede, A. & Wynn, M. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information Systems*.

[28] Van der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer.

[29] Van der Aalst, W. (2015). Extracting Event Data from Databases to Unleash Process Mining. *Management for Professionals*.

[30] Van der Aalst, W., Adriansyah, A., de Medeiros, A., Arcieri, F., Baier, T., Blickle, T., Bose, J., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., ... Wynn, M. (2012). Process Mining Manifesto. *Business Process Management Workshops*.

[31] Van der Aalst, W., van Dongen, B., Günther, C., Mans, R., de Medeiros, A., Rozinat, A., Rubin, V., Song, M., Verbeek, H. & Weijters, A. (2007). ProM 4.0: Comprehensive Support for Real Process Analysis. *Petri Nets and Other Models of Concurrency – ICATPN 2007*.

[32] Van der Aalst, W., Weijters, T. & Maruster, L. (2004). Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*.

[33] Vanbrabant, L., Martin, N., Ramaekers, K. & Braekers, K. (2019). Quality of input data in emergency department simulations: Framework and assessment techniques. *Simulation Modelling Practice and Theory*.

[34] Wang, J., Song, S., Zhu, X. & Lin, X. (2013). Efficient recovery of missing events. *Proceedings of the VLDB Endowment*.

[35] Wang, R. & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*.

[36] Wynn, M. & Sadiq, S. (2019). Responsible Process Mining - A Data Quality Perspective. *Lecture Notes in Computer Science*.