



MSc Computer Science
Thesis

A Data Quality Assessment of the Suspicious Transactions [Public Version]

Ruben Hessels

First supervisor:

dr. ir. Maurice van Keulen

External supervisor:

Elleke van den Brink MSc

Second examiner:

dr. Faizan Ahmed

November, 2023

Department of Computer Science,
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

Contents

1	Introduction	1
1.1	Chain of Events	2
1.2	Problem Statement	4
1.3	Organization	5
1.4	Outline	5
2	Background	6
2.1	Money Laundering and Terrorism Financing Prevention Act	6
2.1.1	What is Money Laundering?	6
2.1.2	What is Terrorism Financing?	7
3	Related Work	8
3.1	Data Quality Methodologies	8
3.1.1	Phase (ii)	9
3.1.2	Phase (iii)	9
3.2	Police	10
3.3	Summary	11
4	Selecting a Data Quality Assessment Methodology	12
4.1	Audit Methodologies	12
4.2	Phase (i) and Phase (ii)	12
4.3	Phase (iii)	13
4.4	Conclusion	15
5	Data Exploration	16
5.1	Dataset Structure	16
5.2	Initial Data Quality Exploration	17
6	Methodology	19
6.1	Selecting Quality Dimensions and Metrics	19
6.2	Measuring Quality Dimensions	19
6.2.1	Completeness	20
6.2.2	Consistent Representation	21
6.2.3	Free-of-error	21
6.2.4	Reputation	22
6.2.5	Timeliness	22
6.2.6	Subjective Quality Dimensions	22
6.3	Measuring Description Quality (Value-added)	23
6.3.1	Length-based	23
6.3.2	Theme-based	24

6.3.3	Labeling Procedure	24
6.3.4	User Goals and Evaluation Metrics	25
6.3.5	Target Value and Threshold	26
7	Results and Discussion	27
7.1	Completeness	27
7.2	Concise Representation	28
7.3	Consistent Representation	29
7.4	Ease of Manipulation	30
7.5	Free-of-error	31
7.6	Record Quality (Value-added)	32
7.6.1	Goal 1: Remove 90% of the Negative Cases	32
7.6.2	Goal 2: Find 90% of the Positive Cases	35
7.6.3	Combining Length and Themes	36
7.6.4	Evaluating Label Quality	37
7.6.5	Data Quality of Records	39
7.6.6	Summary	39
7.7	Understandability	40
7.8	Objectivity	41
7.9	Believability	41
7.10	Reputation	42
7.11	Timeliness	45
8	Application of Research	47
8.1	Dashboard and Record List	47
8.1.1	Dashboard Settings	47
8.1.2	Record List Features	47
8.1.3	User Experience	48
8.2	Hotspots and Themes	49
9	Implications and Prospects	50
9.1	Recommendations	50
9.1.1	FIU	50
9.1.2	Police	51
9.2	Future Steps	52
10	Conclusion	55
A	Reporting Entities	58
B	Dataset Structure	59
C	Dimension Matrix	60
D	Search Terms	62
E	Labeling Criteria	63
F	Completeness	64
G	Reputation	65

H	Threshold Evaluation	68
I	Effectiveness of Social Network Analysis	72
I.1	Goal and Approach	72
I.2	STs as Graphs	72
I.3	Results and Discussion	73
I.3.1	Small Subgraphs	73
I.3.2	Ego Network	74
I.3.3	Groups	75
I.3.4	Semantic Duplicates	75
I.4	Summary	75
	List of Abbreviations	76

Abstract

In recent years, an explosive surge in suspicious transactions has been observed, overwhelming several instances such as the police. In particular, the Regionaal Informatieknooppunt Financieel Economische Criminaliteit (RIK FinEC)-ondermijning, a department within the police that analyzes these transactions on a monthly basis, experiences substantial challenges. They are [REDACTED]

[REDACTED] There is a high demand for new insights about the suspicious transactions to not only uncover data limitations, but also potential improvements in the current approach. This research addresses this issue by conducting a comprehensive data quality assessment of the suspicious transactions dataset. Besides enlightening data quality findings, two novel data quality measures are proposed and implemented as an improvement to the current analytical strategy. On top of that, the research outlines concrete recommendations to both the Financial Intelligence Unit (FIU) and the Dutch police.

Keywords: suspicious transactions, data quality assessment, police

Chapter 1

Introduction

In 2021, the **Financial Intelligence Unit (FIU)** reached a new record by receiving over a million **Unusual Transactions (UTs)**, which, among other irregularities, includes inexplicably large transactions, or transactions to a high-risk country or war zone [24, 25]. Both of which are often associated with money laundering or the financing of terrorism. By law, entities, such as banks and payment service providers, are required to report UTs of their customers to the FIU based on objective and subjective indicators. The blue line in **Figure 1.1** shows an explosive increase in UTs over recent years (a 806% increase with respect to 2012). The FIU considers several criteria for marking UTs as suspicious. The UTs satisfying these criteria are called **Suspicious Transactions (STs)**. An increase in STs can be observed from the red line in **Figure 1.1**, illustrating a nearly 286% increase in a period of 10 years (2012-2022). Due to the overwhelming and severely increased amount of incoming UTs and STs, the police has been challenged with the task of capturing all of the high-potential cases. Over the last three years (2020-2023),

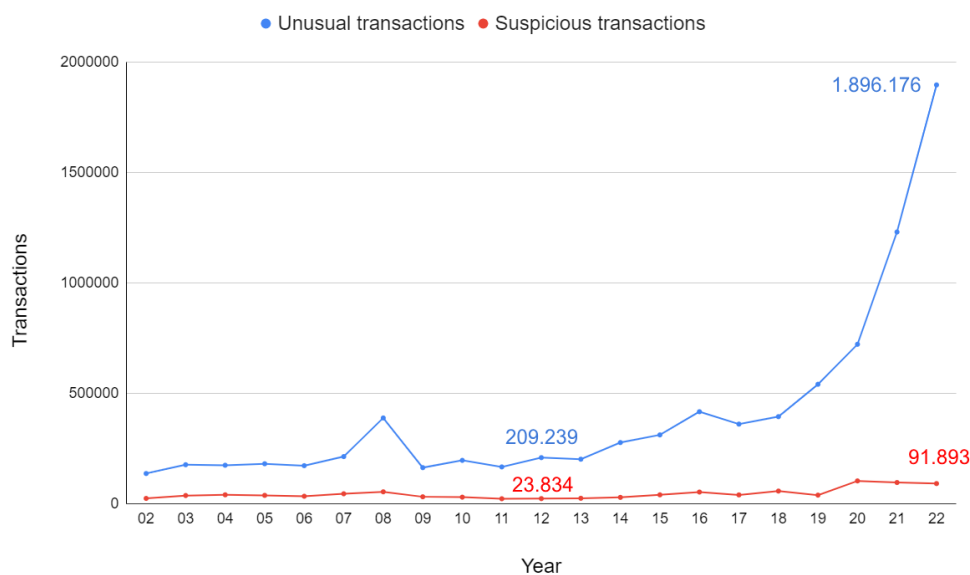


FIGURE 1.1: The amount of unusual and suspicious transactions between 2002 and 2022 [24, 25, 26].

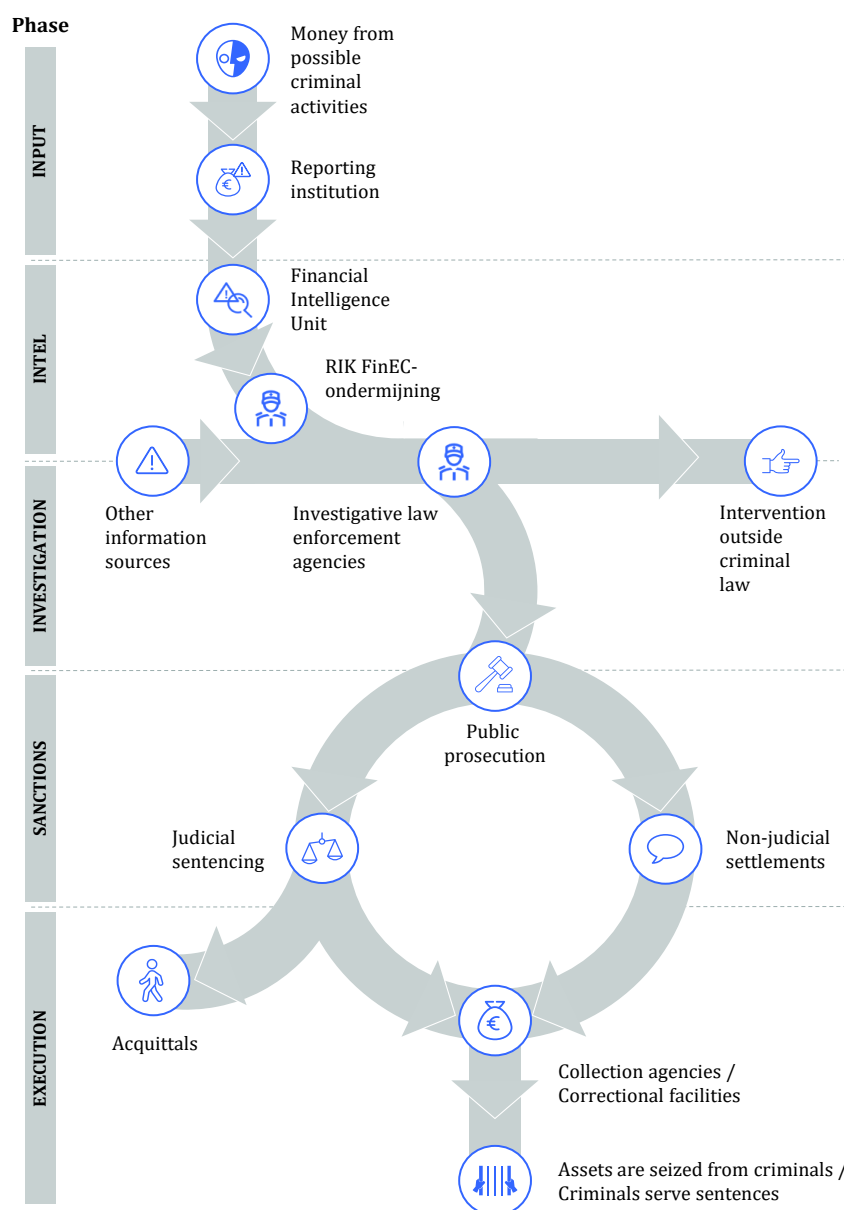


FIGURE 1.2: The full chain of events. The original figure [16] has been translated to English. Furthermore, the RIK FinEC-ondermijning has been included in the chain. Note that RIK FinEC-ondermijning is not necessarily the first to receive STs from the FIU as could be misinterpreted from the figure. In fact, the RIK FinEC-ondermijning is part of the investigative law enforcement agencies as well. However in the context of this research, they are positioned in the chain as illustrated.

3. Recency: [REDACTED]
4. Transaction type: [REDACTED]
5. Suspected offense: [REDACTED]

Both sub-research questions are strongly connected to each other. In fact, the second question is part of the first question in a way that it evaluates a data quality metric that is applied in the first question.

1.3 Organization

This research is carried out at the department of Analysis and Research ([A&O](#)) of the Dutch police unit Midden-Nederland. The client is the [RIK FinEC](#)-ondermijning Midden-Nederland.

1.4 Outline

The outline of this thesis is as follows. First, a small amount of background is discussed in [Chapter 2 Background](#). Second, related literature to this research is reviewed in [Chapter 3 Related Work](#). Third, a data quality assessment methodology is selected in [Chapter 4 Selecting a Data Quality Assessment Methodology](#). Fourth, some initial exploration of the suspicious transaction dataset is conducted in [Chapter 5 Data Exploration](#). Fifth, the research methodology is provided in [Chapter 6 Methodology](#). Sixth, the results and discussion is presented in [Chapter 7 Results and Discussion](#). Seventh, applications of this research are discussed in [Chapter 8 Application of Research](#). Eighth, the implications and prospects of this research are proposed in [Chapter 9 Implications and Prospects](#). Finally, a concise conclusion of this research is given in [Chapter 10 Conclusion](#).

Chapter 2

Background

This chapter offers background information related to the domain of the [RIK FinEC-ondermijning](#).

2.1 Money Laundering and Terrorism Financing Prevention Act

The Money Laundering and Terrorism Financing Prevention Act, or rather [Wet ter voorkoming van witwassen en financieren van terrorisme \(Wwft\)](#), provides a comprehensive set of measures to prevent the use of the financial system for money laundering or terrorist financing [2, 23]. In a general sense, the Wwft is targeted at subversive crime. Subversive crime is represented by an underworld and legitimate business that are becoming increasingly intertwined [14].

2.1.1 What is Money Laundering?

The FIU defines money laundering as follows:

Carrying out acts, or having others carry out acts through which capital gains that have been concealed from the authorities acquire ostensibly legitimate origins. The aim of money laundering is to obscure the provenance of money. [20]

In other words, it is the process of concealing the origins of illegally acquired money from criminal activities such that it appears to have been legitimately earned. In the Netherlands, an *all-crimes approach* is adopted, meaning that all crimes that generate illegally acquired money can be considered a predicate offense for money laundering. Consequently, transactions relating to these crimes are viewed as *unusual* in light of the [Wwft](#) [20].


Money acquired in the *criminal* economy is infiltrated into the *official* economy. The money laundering process always involves three stages [17]:



1. Placement: the illegally acquired money is infiltrated into the economic system.
2. Layering: the infiltrated money is moved around through numerous transactions in an attempt to conceal the origin.
3. Integration: the money is invested (e.g. stocks, real estate).

2.1.2 What is Terrorism Financing?

The ultimate goal of terrorism financing is to provide means, such as money, to support terrorist activities. In contrast to money laundering, terrorism financing could include money that has been legitimately earned.

Whereas in money laundering the main aim is to obscuring the provenance of money, in terrorism financing it is a matter precisely of obscuring its destination. [22]

It should be noted that the RIK FinEC-ondermijning 

¹ 

Chapter 3

Related Work

In this chapter, 13 [Data Quality Assessment \(DQA\)](#) methodologies are reviewed and summarized in [Section 3.1](#). One of which will be used as the methodology for this research. The selection procedure is conducted in [Chapter 6 Methodology](#). Furthermore, related research within the police is given in [Section 3.2](#). A summary of this chapter can be found in [Section 3.3](#).

3.1 Data Quality Methodologies

Assessing the quality of data is typically done by means of a [Data Quality Assessment \(DQA\)](#) (and improvement) methodology. Choosing the most suitable methodology among the literature is not straightforward. A well-known survey conducted by [Batini et al. \[3\]](#) considered 13 methodologies, for which a systematic and comparative description is provided. This section will summarize the steps presented in these methodologies. A methodology is always structured in three phases:

- Phase (i) State reconstruction: collecting contextual information on organizational processes and services, data collections and related management procedures, quality issues and corresponding costs
- Phase (ii) Measurement/assessment: assessing the quality of data collections along relevant quality dimensions. *Measurement* refers to capturing the value of a set of DQ dimensions, while *assessment* refers to comparing the measurements to reference values to establish a diagnosis of the quality.
- Phase (iii) Improvement: setting up the required steps, strategies and techniques to enhance the DQ level

The 13 methodologies can be categorized into four groups. A *complete* methodology provides support to both the assessment and improvement phases. Additionally, both technical and economic issues are addressed. An *audit* methodology focusses on the assessment phase and provides limited support for the improvement phase. The *operational* methodologies focus on technical issues of both the assessment and improvement phase, but not on the economic issues. Finally, the *economic* methodologies focus on the evaluation of costs. The next subsections will expand upon the steps in Phase (ii) and Phase (iii).

3.1.1 Phase (ii)

This subsection will expand upon Phase (ii).

1. Data analysis: examining the data and performing interviews to reach a complete understanding of the data
2. DQ requirements analysis: surveying the opinion of data users to identify quality issues
3. Identification of critical areas: quantitative assessment of relevant databases and data flows
4. Process modeling: providing a model of the processes producing or updating the data
5. Measurement of quality: selecting the quality dimensions affected by DQ issues and defining corresponding metrics. The measurements can be subjective or objective.

Dimensions

Phase (ii) - step 5, is about selecting the quality dimensions affected by the DQ issues. Quality dimensions are perspectives on the [Data Quality \(DQ\)](#) for which metrics can be defined. In the past, numerous sets of DQ dimensions have been proposed by different authors, in which the definitions of the DQ dimensions have subtle differences as well. There is still no agreement on the standard set of DQ dimensions. However, four basic themes are always considered according to [Batini et al.](#): accuracy (the extent to which data are correct, reliable and certified [29]), completeness (the extent to which data is not missing and is of sufficient breadth and depth for the task at hand [28]), consistency (refers to the violation of semantic rules) and timeliness (the extent to which the age of data is appropriate for the task at hand [28]). The DQ dimensions can be assessed either subjectively or objectively. Dimension-specific metrics are context-dependent and can be computed with something as simple as a ratio. For example, completeness is the ratio of non-empty data points divided by the total number of data points. The chosen methodology in [Chapter 6 Methodology](#) will determine which dimensions and corresponding metrics will be used.

3.1.2 Phase (iii)

This subsection will expand upon Phase (iii).

1. Evaluation of the costs: estimation of direct and indirect costs of data quality
2. Assignment of process responsibilities: identifying process owners and their responsibilities on data production and management activities
3. Assignment of data responsibilities: identifying data owners and their data management responsibilities
4. Identification of the causes of errors: identifying causes of data quality problems
5. Selection of strategies and techniques: identifying all the data improvement strategies and corresponding techniques
6. Design of data improvement solutions: selecting the most effective and efficient strategy and related set of techniques

7. Process control: defining check points in the data production process to monitor quality during execution
8. Process redesign: defining process improvement actions that can improve DQ
9. Improvement management: defining new organizational rules for DQ
10. Improvement monitoring: establishing periodic monitoring activities that provide feedback on the results of the improvement process and enables its dynamic tuning

Strategies and Techniques

In Phase (iii) - step 5, a distinction can be made between two types of strategies: *data-driven* and *process-driven*. Data-driven strategies are focused on modifying the data values directly (e.g. updating old values by refreshing the database), while process-driven strategies aim to redesign the processes that create or modify the data (e.g. constraining the format before storing values into the database). Some data-driven improvement techniques are:

1. Acquisition of new data: improving data by acquiring higher-quality data to replace problematic quality values
2. Standardization: altering the data such that it complies with the standard
3. Record linkage: identifying data representations referring to the same object among multiple tables
4. Data and schema integration: defining a unified view of the data
5. Source trustworthiness: selecting data sources based on their DQ
6. Error localization and correction: identifying and eliminates DQ errors
7. Cost optimization: defining quality improvement actions along a set of dimensions by minimizing costs

Two common process-driven strategies are:

1. Process control: inserting checks and control procedures when data are created or updated
2. Process redesign: redesigning processes such that the causes of poor DQ are removed

3.2 Police

This section summarizes some previous work done within and with the police using the [Suspicious Transactions \(STs\)](#).

Zicht op Ondernijning [27] is a dashboard resulting from a partnership between various local and national authorities, initiated by the Ministry of the Interior and Kingdom Relations (BZK). This dashboard combines Statistics Netherlands (CBS) data with data from other reliable national sources such as the [STs](#). The data shown can never be traced back to individual addresses, persons and companies and is thus not designed for actual police investigations. Instead, it is designed to reveal patterns and insights into local criminal phenomena, which can be used by local authorities.

[REDACTED] de Lignie et al. [6] [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

Goudzand [9] [REDACTED]
[REDACTED] Goudzand [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

3.3 Summary

Data quality is assessed by means of a [Data Quality Assessment \(DQA\)](#) methodology. Several types of DQA methodologies exist, each consisting of three phases. A DQA methodology should be selected based on the desired steps within these phases ([Chapter 4 Selecting a Data Quality Assessment Methodology](#)). Furthermore, no data quality assessments of the suspicious transactions have been conducted at the police, as far as is known.

Chapter 4

Selecting a Data Quality Assessment Methodology

In the previous chapter, [Section 3.1](#) summarized the common steps followed in the 13 [Data Quality Assessment \(DQA\)](#) procedures covered by [Batini et al.](#). The most suitable methodology for this research will be chosen from this selection by evaluating which steps in Phase (ii) and Phase (iii) are most relevant. The most suitable methodology will serve as the foundation of the DQ assessment procedure that will be applied in [Chapter 6 Methodology](#). The conclusion of this chapter is provided in [Section 4.4](#).

4.1 Audit Methodologies

The goal of the first sub research question is to assess the quality of the dataset. Besides an assessment phase, DQ methodologies often consist of an improvement phase as well. However, for this research, the assessment phase is deemed more important than the actual execution of the improvement steps. The reason for this is that applying significant improvements to the source data is a task more fitting to be executed by the FIU than by the police. Therefore, an *audit* methodology would seem very suitable since this mostly focusses on the assessment phase and provides limited support for the improvement phase. Specific improvements may be suggested, but will not necessarily be executed. Considering only the audit methodologies already cuts the total selection in half (i.e. 6 of the 13 methodologies remain). The remaining (abbreviated) audit methodologies include: AIMQ [\[10\]](#), CIHI [\[12\]](#), DQA [\[15\]](#), AMEQ [\[18\]](#), QAFD [\[1\]](#) and IQM [\[8\]](#).

4.2 Phase (i) and Phase (ii)

All methodologies contain a similar Phase (i) in which information is gathered about the stakeholders and data source. However, differences among the audit methodologies arise in Phase (ii). For this research, only step 1 and 5 are considered to be relevant. An overview of the selected Phase (ii) steps is provided in [Table 4.1](#).

According to [Batini et al.](#), all of the audit methodologies support step 1 and all of them, except for CIHI, support step 5. An overview is provided in [Table 4.2](#). IQM shows to be an exact match with the desired Phase (ii) steps (i.e. only step 1 and 5 are supported). DQA and AIMQ match the desired Phase (ii) steps except for one step.

TABLE 4.1: Relevant steps in Phase (ii)

#	Steps	Include	Reason
1	Data analysis	x	Reaching a complete understanding of the data is a critical step.
2	DQ requirements analysis		There will be no survey of the opinion of data users.
3	Identification of critical areas		There is only one database to be examined.
4	Process modeling		The process model would mainly be about processes within the FIU and thus out of scope.
5	Measurement of quality	x	The quality dimensions will be selected and measured.

TABLE 4.2: Audit methodologies and their supported assessment steps. All methodologies support step 1 and only CIHI does not support step 5.

#	Steps	AIMQ	CIHI	DQA	AMEQ	QAFD	IQM
1	Data analysis	x	x	x	x	x	x
2	DQ requirement analysis			x		x	
3	Identification of critical areas	x	x		x	x	
4	Process modeling				x		
5	Measurement of quality	x		x	x	x	x
Suitable		x		x	x	x	x

4.3 Phase (iii)

Phase (iii) is targeted at implementing improvement steps. Although improving the DQ is not the main objective of this research, improvements will be applied when possible. Another purpose of Phase (iii) is not to necessarily to execute the improvement steps, but to compose recommendations to the data owners (the [FIU](#)). The relevant steps in Phase (iii) are step 4 and 5 (refer to [Table 4.3](#)).

TABLE 4.3: Relevant steps in Phase (iii)

#	Steps	Include	Reason
1	Evaluation of the costs		There will be no economic assessment.
2	Assignment of process responsibilities		The focus will not be on the process.
3	Assignment of data responsibilities		The focus will not be on owners and responsibilities.
4	Identification of the causes of errors	x	After measuring the DQ, it is important to address the causes of certain problems.
5	Selection of strategies and techniques		Approaches for improving the data will be explored.
6	Design of data improvement solutions		Data improvement solutions will not be designed.
7	Process control		The focus will not be on the process.
8	Process redesign		The focus will not be on the process.
9	Improvement management		The focus will not be on the organization.
10	Improvement monitoring		The improvement process will not be monitored.

Table 4.4 shows that DQA and AMEQ are the only audit methodologies that provide support for improvement steps. Both DQA and AMEQ show support for the identification of the causes of errors step, but not the selection of strategies and techniques. AMEQ also supports improvement monitoring, but this is not desired in this research. Therefore, DQA (Data Quality Assessment) by Pipino et al. [15] seems to be the best match to the needs of this research. In the previous phase, IQM had a better match with the desired assessment steps than both DQA and AMEQ. However, IQM provides no improvement steps at all and is thus not considered.

TABLE 4.4: Audit methodologies and their supported improvement steps. Note that few steps are supported, since audit methodologies do not typically focus on improvement steps.

#	Steps	AIMQ	DQA	AMEQ	QAFD	IQM
1	Evaluation of the costs					
2	Assignment of process responsibilities					
3	Assignment of data responsibilities					
4	Identification of the causes of errors		x	x		
5	Selection of strategies and techniques					
6	Design of data improvement solutions					
7	Process control					
8	Process redesign					
9	Improvement management					
10	Improvement monitoring			x		
Suitable			x	x		

It should be noted that other methodologies types such as *complete* and *operational* frequently support both of the desired steps in Phase (iii). However, they often focus on

various other steps as well. Consequently, the focus of these methodologies is often on either process-related steps or cost-related steps. This research is focused on neither of those, but rather on audit methodologies.

DQA supports all of the desired steps. In addition, it supports the identification of critical areas in Phase (ii). This step can be omitted since a quantitative assessment of relevant databases and data flows is unnecessary. The assessment will only be for one database. Moreover, there are no relevant data flows present.

4.4 Conclusion

DQA (Data Quality Assessment) by [Pipino et al. \[15\]](#) is the most suitable methodology for this research based on its supported assessment and improvement steps. DQA will be used with a slight variation by removing Phase (ii) - step 3. The included steps can be found in [Table 4.1](#) and [Table 4.3](#).

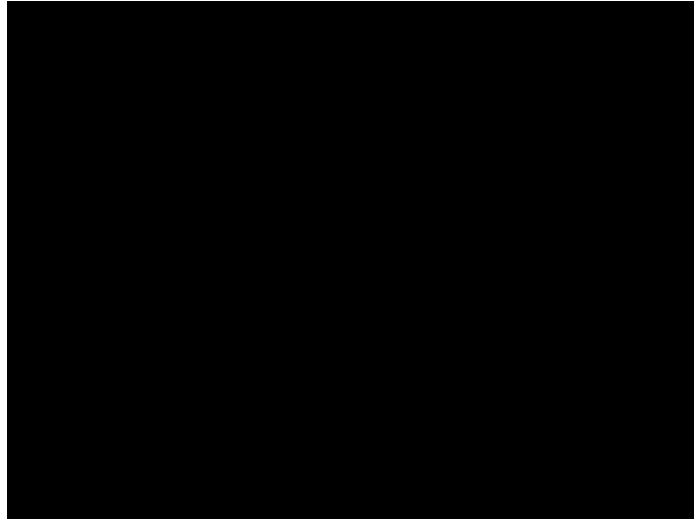


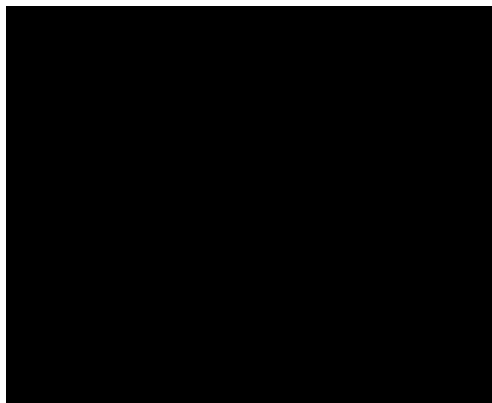
FIGURE 5.1: A visualization of a STs record

5.2 Initial Data Quality Exploration

This section is the result of Phase (ii) - Data analysis (step 1) of the DQA methodology discussed in [Chapter 4 Selecting a Data Quality Assessment Methodology](#).

Note that [Figure 5.1](#) [REDACTED] [Table 5.1](#). The median/50th percentile [REDACTED] In [Appendix I](#), [REDACTED] [REDACTED] [REDACTED] [REDACTED]

TABLE 5.1: The distribution of parties and transactions viewed per record



[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

Chapter 6

Methodology

This chapter proposes a methodology guided by the [Data Quality Assessment \(DQA\)](#) methodology of [Pipino et al.](#) [Chapter 5](#) conducted the *Data analysis* step of Phase (ii), while this chapter proposes the approach for the last remaining step of Phase (ii), which is the *Measurement of quality* step.

6.1 Selecting Quality Dimensions and Metrics

Measuring [Data Quality \(DQ\)](#) is typically achieved by utilizing so-called quality dimensions. The dimensions proposed by DQA are listed in [Table 6.1](#) and contain a general set of dimensions that may be applied. Depending on the data, a selection of relevant dimensions can be extracted. In the case of this research, a selection is made with respect to the [Suspicious Transactions \(STs\)](#). A column in the STs can either be appropriate to be assessed by a dimension or not. This produces a matrix presented in [Appendix C](#). Since the dimensions *Interpretability* and *Understandability* are viewed to be similar, these dimensions are combined¹. Similarly, *Relevancy* and *Value-added* are merged².

In order to quantify the dimensions in [Table 6.1](#), metrics were defined by [Pipino et al.](#). Three functional forms that act as objective DQ metrics were considered: (i) simple ratio: the ratio of the desired outcomes to the total outcomes (ii) min or max operation: the minimum and maximum operation to the dimensions that require aggregation of multiple DQ indicators (iii) weighted average: the average of variables with each an assigned weight (weights are normalized and should sum up to 1). For objective criteria measurements, the simple ratio is only used in this research. However, it is not always possible to quantify a dimension. In this scenario, a subjective measurement is more appropriate.

6.2 Measuring Quality Dimensions

The selected DQ dimensions in [Table 6.1](#) are measured either objectively or subjectively. For objective dimensions, this section formulates approaches using the simple ratio metric. For subjective dimensions, approaches are described that aim to measure the respective dimension in the best way viewed. The subjective dimensions are: *Believability*, *Concise representation*, *Ease of manipulation*, *Objectivity* and *Understandability*. The objective dimensions are: *Completeness*, *Consistent representation*, *Free-of-error*, *Reputation*, *Timeliness* and *Value-added* (will be covered in [Section 6.3](#) separately). Recall that only the

¹Will be referred to as *Understandability* in this research

²Will be referred to as *Value-added* in this research

TABLE 6.1: Quality dimensions [15] and their definitions

Quality dimensions	Definition	Relevant
Accessibility	the extent to which data is available	
Appropriate amount of data	the extent to which the volume of data is appropriate for the task at hand	
Believability	the extent to which data is regarded as true and credible	x
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand	x
Concise representation	the extent to which data is compactly represented	x
Consistent representation	the extent to which data is presented in the same format	x
Ease of manipulation	the extent to which data is easy to manipulate and apply to different tasks	x
Free-of-error	the extent to which data is correct and reliable	x
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial	x
Reputation	the extent to which data is highly regarded in terms of its source or content	x
Security	the extent to which access to data is restricted appropriately to maintain its security	
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand	x
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear	x
& Understandability	& the extent to which data is easily comprehended	x
Relevancy	the extent to which data is applicable and helpful for the task at hand	x
& Value-added	& the extent to which data is beneficial and provides advantages from its use	x

STs columns marked in [Appendix C](#) are measured and that different dimensions cover different aspects of the data. As a result, each column covered by the dimension has a certain computed score. [REDACTED]

6.2.1 Completeness

Completeness measures the number of non-empty cells in a column. However, a non-empty cell is not guaranteed to hold meaningful data. As of such, certain “noise” characters are regarded as “incomplete” data. [REDACTED]

[REDACTED] However, it is acknowledged that this approach does not provide a perfect coverage of the actual noise in the dataset.

In some cases, a cell in a column is expected to be empty. [REDACTED]

[REDACTED] (KvK) [REDACTED]

6.2.2 Consistent Representation

In order to measure consistency in the format among the columns in the data, the desired formats should first be defined. Not all columns adhere to a certain format and can thus not be measured. However, columns that do contain structured data can be measured using [Regular Expression \(RegEx\)](#), which is a way to describe patterns within strings of characters. It is a powerful method to perform matching tasks, or rather, to determine whether the pattern or format of the data is correct. Note that formats may differ internationally (e.g. zip codes). Therefore, some of the RegExs are defined for the Netherlands only. The amount of consistent data divided by the total amount of “complete” data (refer to [subsection 6.2.1](#)) represents the final consistency score.

In total, [REDACTED]

[REDACTED] ([Kamer van Koophandel \(KvK\)](#)) [REDACTED]

[REDACTED] it is checked whether they follow the [International Bank Account Number \(IBAN\)](#) format correctly. The bank account address should start with the country code (i.e. NL), two control digits, four bank identifier letters and the account number (ten digits).

6.2.3 Free-of-error

Quantifying the amount of errors in the data is a challenging task and requires some form of ground truth. Since this is not always available, only a small selection of data is verified in this research. [REDACTED]

[REDACTED] [schwifty \[7\]](#) Python library.

6.2.4 Reputation

Reputation is defined as the extent to which data is highly regarded in terms of its source or content. Since the sources of the data are the reporting entities, the reputation dimension is measured by determining their reputation. This is achieved by using the measures proposed in [Section 6.3](#). Each record can then be associated to a data quality score. Since reporting entities are linked to records, a connection can be established between the data quality of a record and the reporting entity. In this way, a reputation score is defined in terms of the data quality.

6.2.5 Timeliness

The age of the data is an important data quality aspects and reveals to what extent the data is sufficiently up-to-date. [REDACTED]

6.2.6 Subjective Quality Dimensions

The previous subsections discussed the approaches for measuring the objective dimensions. This subsection proposes the approaches for measuring the subjective quality dimensions. In general, the procedure of measuring subjective dimensions is less pronounced and thus described in a shorter fashion.

Concise Representation

Conciseness is measured subjectively by investigating the presence of redundant data. Since the dataset consists of two dimensions (columns and rows), both dimensions should be inspected for redundancies. This is achieved by inspecting whether certain columns can be omitted without losing any unique information. Every row in the dataset represents a party. When rows are redundant, this would imply that duplicate parties are present. Both of the aforementioned scenarios are investigated and may raise awareness of the current effectiveness of data storage.

Ease of Manipulation

All columns are subjectively ranked into three categories based on how easy these columns are to manipulate and apply to different tasks: *Easy*, *Medium* and *Hard*.

Understandability

This combined dimension (refer to [Section 6.1](#)) is subjectively measured by manually judging whether each column can be easily understood directly without requiring additional knowledge.

Objectivity

This dimension evaluates whether certain columns might be prone to subjectivity by assessing what data the reporting entity was required to enter.

Believability

This dimension criticizes what data can be regarded as true and credible, which is a concern since the data originates from various reporting entities.

6.3 Measuring Description Quality (Value-added)

The previous section proposed approaches for measuring the selected quality dimensions. An approach for measuring the *Value-added* dimension is yet to be proposed. It is discussed in this separate section since it requires a relatively more elaborate methodology than the other dimensions.

First, it is necessary to define how this quality dimension will be interpreted in the context of this research. [Appendix C](#) shows that *Value-added* only covers the [REDACTED]. It is thus solely focussed on the [REDACTED]. The assessment of this column can be performed in two ways. [REDACTED]

[REDACTED] ([subsection 6.3.1](#)). The idea for using this indicator originates from the fact that [REDACTED]. In addition, [Boekholt \[4\]](#) found that length is an important indicator for extracting high-potential STs records when training machine learning classifiers. The second indicator is the number of topic-related keywords ([subsection 6.3.2](#)) in all transaction descriptions of a record. The latter is directly related to the second research sub question, asking how theme-based measures can be used to measure the quality of transaction records.

6.3.1 Length-based

Since the record quality³ is measured and not the transaction quality, the transaction descriptions of a record are concatenated into one string. This will be referred to as the *merged description* from now on.

The first method assumes that the longer the description is (character-wise), the higher its text quality. Additionally, a variation, in which the merged description length is divided by the number of transactions (equal to the average transaction length), is applied as well.

For the purpose of studying the effect of the number of transactions on the total description length, [REDACTED]

³ [REDACTED]

Figure 6.1, in which the blue line represents the average description length for each number of transactions and the orange line the trend in the blue plot.

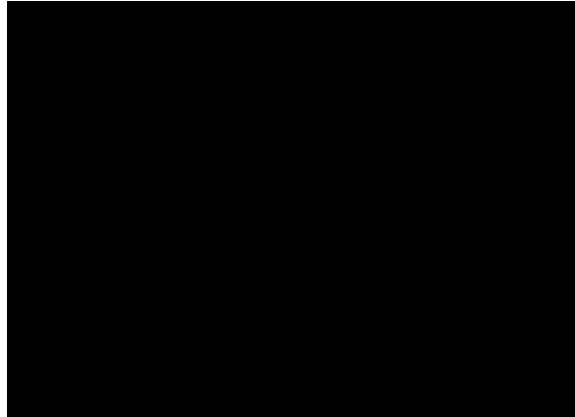


FIGURE 6.1: The relationship between the description length and number of transactions.

6.3.2 Theme-based

The second one is based on the idea that elaborate descriptions have a higher chance of holding interesting topic-related keywords. Using both methods, the records in the dataset can be split into two subsets: low DQ and high DQ. This split is performed by a certain threshold, which for the length-based method would be the number of characters and for the theme-based method the number of keywords present in the description.

The core of the topic-related keywords consists of keywords originating from the 2022 [Anti Money Laundering Centre \(AMLC\)](#) annual report [5]. The report covers themes . It is unknown whether the AMLC keywords appropriately cover each theme since they have not been evaluated.

Additionally, a variation is created in which only “interesting” themes, and thus interesting keywords, are included. Here, “interesting” is defined based on what themes the RIK FinEC-ondermijning finds interesting (based on their experience). An overview of all the AMLC themes/keywords, interesting keywords and newly added themes/keywords can be found in [Appendix D](#). The newly added themes (below horizontal line) and keywords (boldfaced) were extracted from the data based on manual observations.

6.3.3 Labeling Procedure

In order to evaluate these approaches, a ground truth is required. More specifically, labels that indicate whether a record is of low or high quality. But first, certain operations should be applied to the data to make them ready for labeling such that it resembles the dataset the RIK FinEC-ondermijning uses when analyzing the records. The procedure is as follows:

1. Remove transactions

2. Remove transactions unrelated to the Midden-Nederland region.
3. Group all the [REDACTED]

A 1 to 5 label score is assigned to a record. Four people, two of which of the RIK FinEC-ondermijning, referred to as the experts, and two students, referred to as the non-experts, were involved in the labeling process. Consequently, a bias in determining the scores might occur. For this reason, some labeling criteria have been specified in [Appendix E](#) to ensure consistency. Generally, the labels 1 and 2 represent low-quality records, while 3 to 5 represent high-quality records.

6.3.4 User Goals and Evaluation Metrics

In order to evaluate both the length-based and theme-based approach, various evaluation metrics are suitable depending on the goal of their usage. In practice, two goals are realistic using these methods: *Goal 1*, filter out as many low DQ records; *Goal 2*, capture as many high DQ records

Goal 1 may be suitable if the user aims to eliminate as many meaningless records as possible. A danger of this approach is that it eliminates a portion of the high DQ records as well, leaving a relatively small portion of high DQ records. Nevertheless, if the user is determined to read a small selection of records anyway, then this consequence is less concerning. *Goal 2* may be suitable if the user aims to capture a complete view of all the high DQ records. A side-effect of this goal is that many low DQ records might be included in the high DQ selection, leaving a substantial number of records to read.

[REDACTED]

so both goals will be taken into account during this research.

For *Goal 1*, false positives (i.e. low DQ records classified as high DQ) must be minimized. Precision ([Equation 6.2](#)) and specificity ([Equation 6.1](#)) are the evaluation metrics of interest. While precision focuses on the performance of the positive class (high DQ), specificity is targeted at the negative class (low DQ). Therefore, specificity is more interesting for this goal as it reveals how effective the negative class is filtered out.

$$specificity = \frac{TN}{TN + FP} \tag{6.1}$$

$$precision = \frac{TP}{TP + FP} \tag{6.2}$$

Goal 2 aims to maximize true positives (i.e. high DQ records predicted as high DQ) and minimize false negatives (i.e. high DQ records predicted as low DQ) at the cost of allowing some false positives. Recall ([Equation 6.3](#)), can be considered the opposite of specificity, is well-suited for this objective as it focuses on the correct prediction of positive cases (high DQ). Furthermore, F-score ([Equation 6.4](#), in which $\beta = 1$ is the harmonic mean

of precision and recall) might be a relevant additional metric. In addition to recall, not only false negatives are considered in this case, but also false positives due to including precision. The effect of allowing false positives in recall can then be observed in this new F-score value (i.e. if the model performs well on recall, but allows many false positives, the F-score will be low compared to recall).

$$recall = \frac{TP}{TP + FN} \tag{6.3}$$

$$F_{\beta} = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \tag{6.4}$$

In short, specificity is used for goal 1 and recall for goal 2. F-score is included as well for a broader view of performance.

6.3.5 Target Value and Threshold

The target value is a performance percentage that the user specifies when selecting the aforementioned length-based and theme-based methods. The target value specified by the RIK FinEC-ondermijning is 90% (note that for F-score this may not be necessarily achieved). For *Goal 1*, this translates to the objective that 90% of the records filtered out should in fact be of low DQ, which directly implies that a 10% error margin is allowed (i.e. high DQ records classified as low DQ). Conversely for *Goal 2*, 90% of the records classified as high DQ should be of high DQ, or rather 10% of the predicted high DQ records are allowed to be low DQ. Note that a target value of 100% is not desirable as it always comes at a large cost of the performance of another metric.

The threshold variable determines the performance of the evaluation metrics. Shifting this threshold causes records to be differently classified. For example, moving the length threshold upwards means that less records are classified as high DQ and thus changes the evaluation results. When aiming for a certain target value, the optimal threshold must be found. This can be achieved by computing all evaluation metrics for all possible thresholds. Then, the optimal threshold can be determined by finding the threshold performing closest to the desired target value.

7.3 Consistent Representation

Consistent representation is defined as to which extent the data is represented in the same format. [Table 7.1](#) shows the results of the consistency dimension. The *Consistent* column indicates the number of cells representing the same format. The *Total* column represents the number of non-empty and non-noisy cells, or rather, the cells containing data (extracted from measuring completeness in [Section 7.1](#)). The *Consistency* column shows the percentage of cells that follows representing the same format.


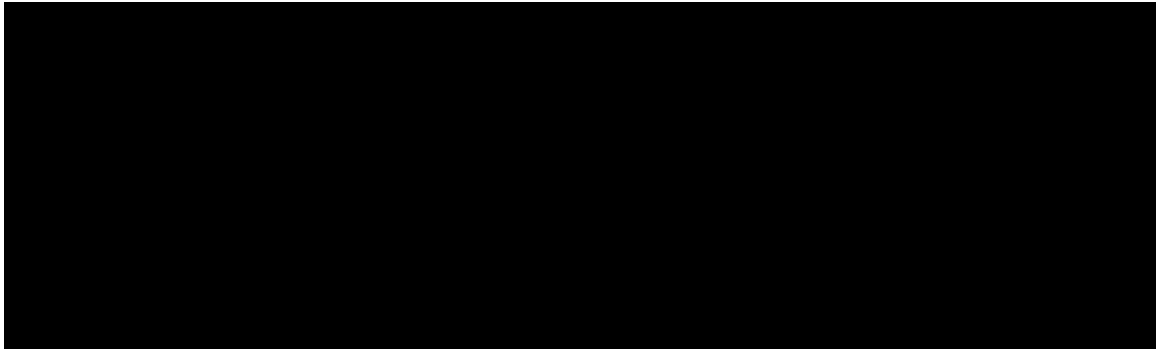
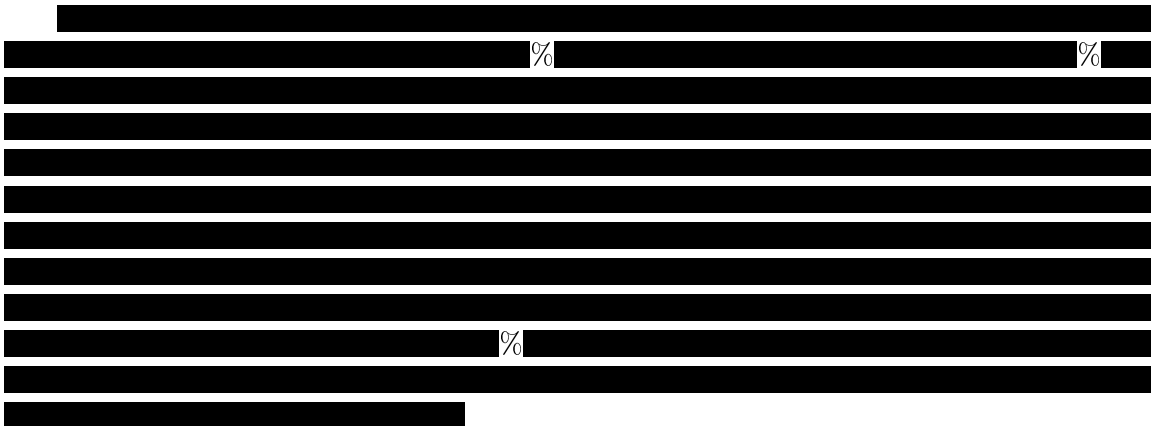
A table with multiple rows and columns, almost entirely redacted with black bars. Only a few cells are visible, including a percentage sign (%) in one of the lower rows.

TABLE 7.1: Consistency results

A large rectangular area that has been completely redacted with a solid black fill, obscuring all content underneath.A table with multiple rows and columns, almost entirely redacted with black bars. Only a few cells are visible, including percentage signs (%) in two of the lower rows.

[Table 7.2](#) presents the cases that were tested to verify the correctness of the respective RegEx. It provides an overview of which cases are included and excluded by following the desired format.

TABLE 7.2: Test results *Consistent representation*



7.4 Ease of Manipulation

Ease of manipulation is defined as the extent to which data is easy to manipulate and apply to different tasks. This dimension is connected to the dimensions consistency and free-of-error (next section). Because when these are low, the data becomes less usable. Most columns are not covered by the consistency and free-of-error dimensions. Therefore, most columns are subjectively ranked.

Table 7.3

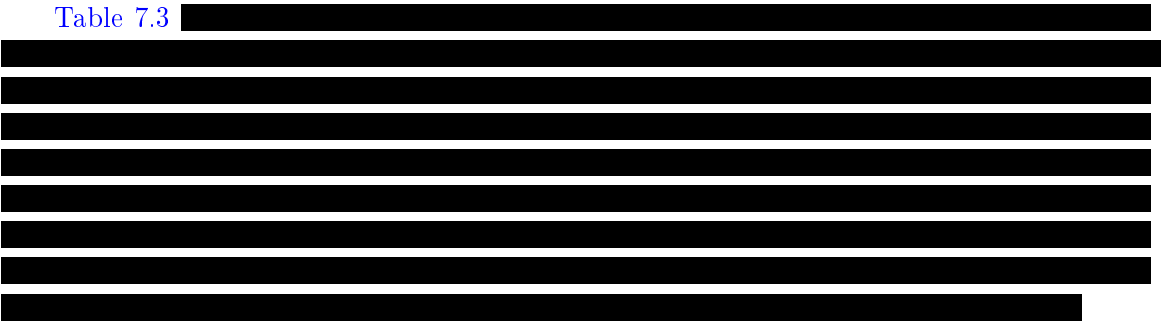
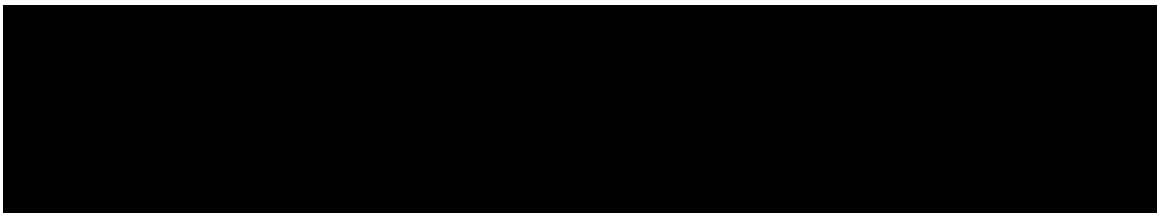


TABLE 7.3: Three levels for ease of manipulation



International
Bank Account Number (IBAN)

[REDACTED]

7.5 Free-of-error

Free-of-error measures the extent to which data is correct and reliable. [Table 7.4](#) presents the results for the free-of-error DQ dimension.

TABLE 7.4: Free-of-error results

[REDACTED]

[REDACTED]

[REDACTED] [Table 7.5](#) gives an overview of how various cases were included and excluded.

TABLE 7.5: Free-of-error test results



7.6 Record Quality (Value-added)

Value-added is defined as the extent to which data is beneficial and provides advantages from its use. This dimension has been applied to transaction descriptions (*Beschrijving transactie*) only. In this research, value-added is interpreted as the extent to which data can be utilized for approximating the quality of a record. Note that, the descriptions of all transactions in a record have been concatenated into one single description (referred to as the *merged* description) in order to assess the quality on a record level rather than on a transaction level.

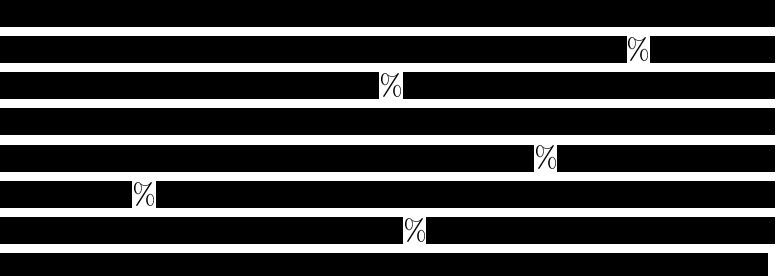
Two length-based and two theme-based methods are explored in this section. Length-based measures use the length of the merged description, in terms of characters, to approximate the quality of a record. Theme-based measures use the number of topic-related keywords extracted from the merged description to determine the quality of a record. Using these methods, records can be classified as either low DQ or high DQ based on a specified threshold. Records scoring above the threshold are considered to be of high DQ and vice versa.

The evaluation of these methods depends on certain user goals (refer to goal 1 and 2 specified in [Section 6.3](#)). Goal 1 entailed filtering out as many low DQ records as possible, while minimizing false negatives. Conversely, goal 2 consisted of finding as many high DQ records as possible, while minimizing false positives. For goal 1, specificity was selected, while recall and F-score were for goal 2. On top of that, a target value of 90% was selected for both goals by the RIK FinEC-ondermijning. The evaluation results will reveal which thresholds should be adhered to with respect to different target values. Finally, a summary of this section is provided in [subsection 7.6.6](#).

7.6.1 Goal 1: Remove 90% of the Negative Cases

Length

[Figure 7.1](#) shows 

[Table H.1](#) in [Appendix H](#) 

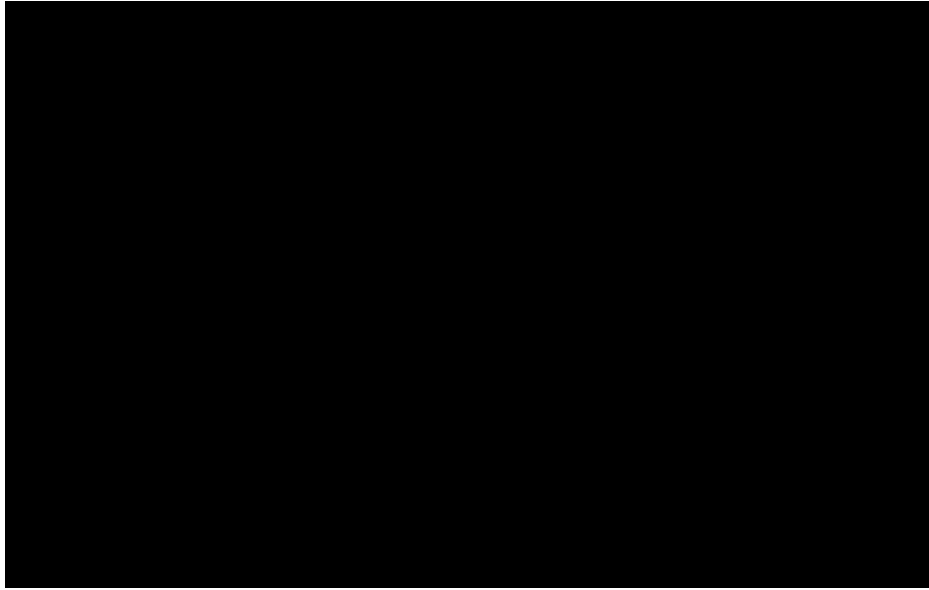


FIGURE 7.1: Performance of evaluation metrics for length

Normalized Length

Similar to [Figure 7.1](#), [Figure 7.2](#) 



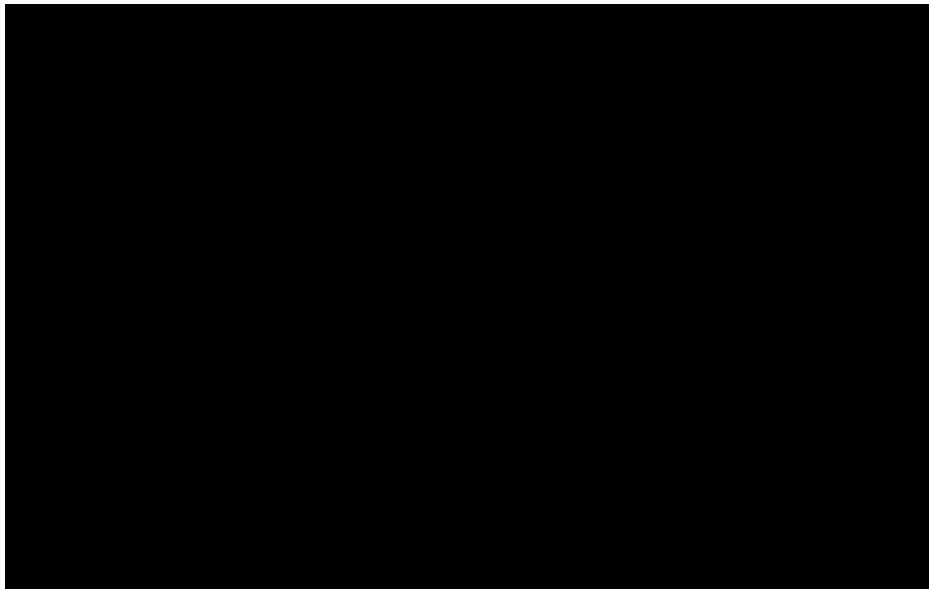





FIGURE 7.2: Performance of evaluation metrics for normalized length

[Table H.2](#) %  (similar to
length in [Table H.1](#)), % 
% % % 
% % 

[REDACTED]

Keywords

Figure 7.3 presents the performance plot for keywords. [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] %

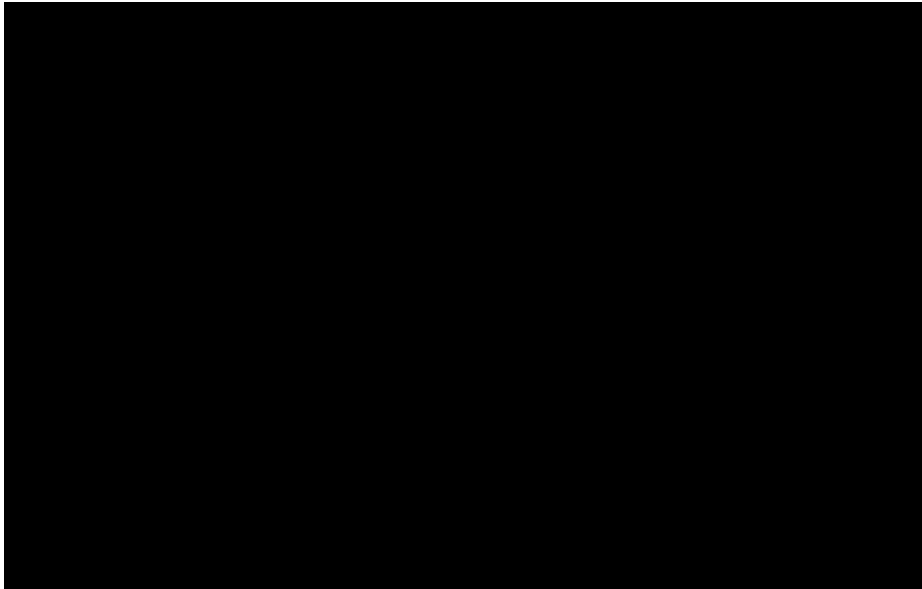


FIGURE 7.3: Performance of evaluation metrics for keywords

Comparing this recall percentage to the length-based methods reveals that length-based methods easily outperform the number of keywords (Table H.3). When the target score is focused on specificity, the associated recall scores are consistently higher. In fact, this is true for all different target scores.

Interesting Keywords

Figure 7.4 [REDACTED] % [REDACTED] (refer to Table H.4). [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] % [REDACTED] %

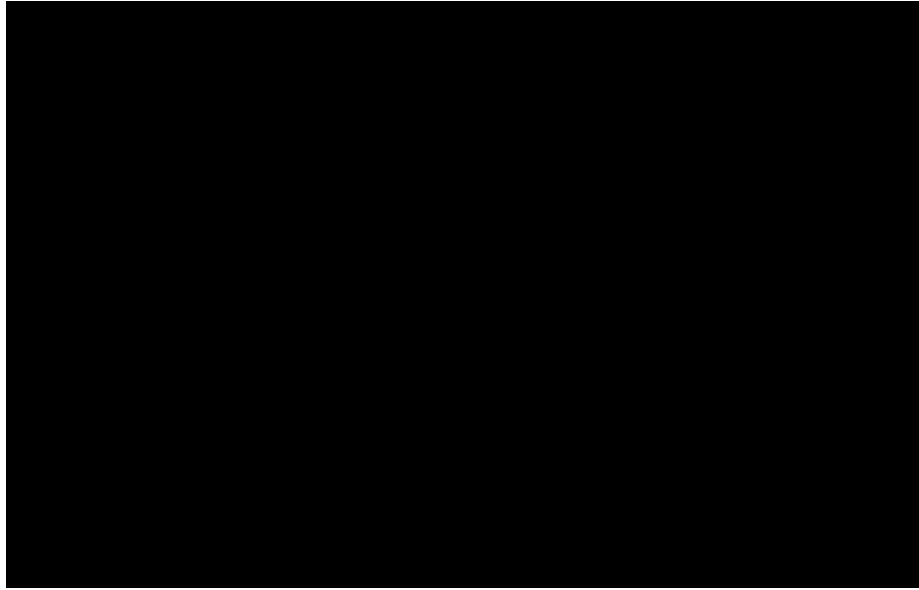


FIGURE 7.4: Performance of evaluation metrics for interesting keywords

Furthermore, similarly to the keywords method, the length-based methods consistently outperform the theme-based methods.

7.6.2 Goal 2: Find 90% of the Positive Cases

Length

Figure 7.1 [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

Table H.1 in Appendix H [REDACTED]

[REDACTED] %

[REDACTED] %

[REDACTED]

[REDACTED] %

[REDACTED] %

Normalized Length

Figure 7.2 [REDACTED]

[REDACTED] %

[REDACTED] % (refer to Table H.2). [REDACTED] %

[REDACTED]

[REDACTED]

TABLE 7.6: Recall performance of combined indicators compared to the length indicator




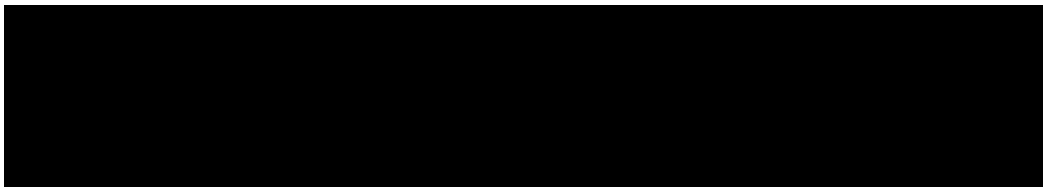
When a similar analysis is conducted for [Table 7.7](#), it is shown that different targeted specificity scores do not outperform the original length measure. 



TABLE 7.7: Specificity performance of combined indicators compared to the length indicator



To conclude, the optimal combination of length and interesting keywords does not improve performance compared to using only length.

7.6.4 Evaluating Label Quality


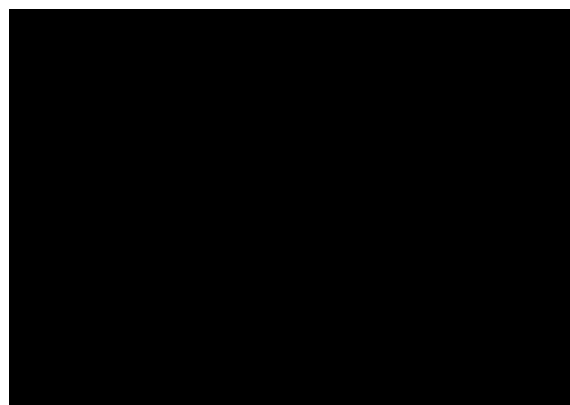
The labels used in this section about the description quality can be distinguished into two sets (refer to [Table 7.8](#)). The first (expert) set has been created by the RIK FinEC-ondermijning and  Furthermore, a second (non-expert) set has been created which contains randomly sampled records. Since the first set is labeled by the experts, it can be compared to the non-expert set in order to validate the reliability of these non-expert labels.

TABLE 7.8: Distribution of labels



The mean and standard deviation in Table 7.8 reveal certain characteristics in the labeling behavior for the randomly sampled two sets. [REDACTED]

[REDACTED]

Furthermore, it can be observed that [REDACTED] This could either mean that [REDACTED] The former begs the question whether the [REDACTED]

A common approach for assessing the sample size is by applying statistical methods. The sample size of the expert set will be assessed by using the computed sample mean and standard deviation in Table 7.8. This procedure will only be performed using the sample statistics from the expert set since these are more reliable and closest to the ground truth.

Given are the sample size ([REDACTED]) of the expert set and the targeted confidence level ($c = 0.95$) of 95%. The degrees of freedom (df) is calculated as: [REDACTED]

A t-test is appropriate here because the population standard deviation (σ) of the total dataset ([REDACTED]) is unknown. Furthermore, the population mean (μ) is estimated with the sample mean $\bar{\mu} = [REDACTED]$ and sample standard deviation $\bar{\sigma} = [REDACTED]$. The corresponding critical value t_α for an area of $\alpha = 1 - c = 0.05$ and $df = [REDACTED]$ for a two-tailed t-test is approximately [REDACTED].

The margin of error (ME) is computed by:

$$ME = t_\alpha * \left(\frac{\bar{\sigma}}{\sqrt{n}}\right)$$

$$ME = [REDACTED] * \left(\frac{[REDACTED]}{\sqrt{[REDACTED]}}\right) \approx [REDACTED]$$

The confidence interval (CI) can then be constructed by:

$$CI = \bar{\mu} \pm ME$$

$$CI \approx [REDACTED] \pm [REDACTED]$$

Which results in an approximate range of $CI \approx [REDACTED]$, in which there is a 95% confidence that the actual population mean (μ) falls within this interval. Since the sample mean $\bar{\mu}$ [REDACTED]

[REDACTED] In hindsight, only the expert labels could have been used for evaluation.

To conclude, differences in labeling behavior can be observed between the expert set and non-expert set. Although they are not necessarily pronounced, they should be considered a

contributing factor that affects the evaluation quality of the length-based and theme-based methods.

7.6.5 Data Quality of Records

The previous subsections discussed and evaluated the results of the proposed length-based and theme-based measures. The initial purpose of these methods was to measure the data quality of merged descriptions. This subsection reveals how the methods perform on the total dataset with their optimal threshold (target value 90%).

Table 7.9 [redacted]%. However, the evaluation in subsection 7.6.1 [redacted]

TABLE 7.9: Goal 1: The percentage of low DQ records that is removed with the optimal threshold for a 90% target value

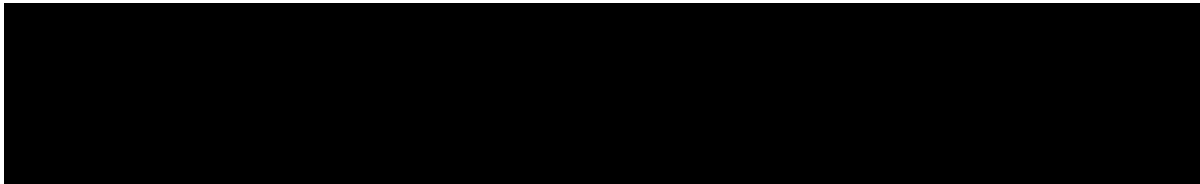
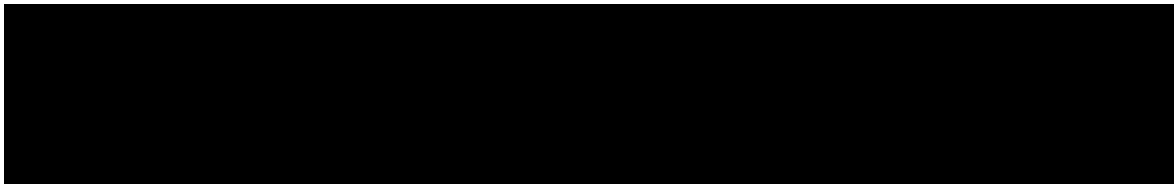


Table 7.10 [redacted]%

TABLE 7.10: Goal 2: The percentage of high DQ records that is selected with the optimal threshold for a 90% target value



Overall, [redacted] Implementing this as an additional filter step for the monthly datasets might accelerate certain procedures of the RIK FinEC-ondermijning.

7.6.6 Summary

To summarize this dimension, length-based methods outperform theme-based methods. [redacted]

7.7 Understandability

Interpretability is defined as the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear. This quality dimension was combined with understandability (refer to [Section 6.1](#)), which is defined as the extent to which data is easily comprehended.

[REDACTED]

[REDACTED] in [Section 7.6](#). [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

7.8 Objectivity

Objectivity is defined as the extent to which data is unbiased, unprejudiced, and impartial.

[REDACTED]

[REDACTED]

[REDACTED] Therefore, it is not possible to determine the objectivity of this column.

7.9 Believability

Believability is defined as the extent to which data is regarded as true and credible. [REDACTED]

[REDACTED]

7.10 Reputation

Reputation is defined as the extent to which data is highly regarded in terms of its source or content. Since the sources of the data are the reporting entities, the reputation dimension is measured by determining their reputation. This can be achieved by applying the evaluated length-based and theme-based measures of [Section 7.6](#) to each record, such that each record is associated with a data quality score. Because reporting entities are linked to records, a connection can be established between the data quality of a record and the reporting entity. In other words, the average record quality a reporting entity delivers, is revealed.

Records consisting of multiple reporting entities are not taken into account in this dimension, since one entity may contribute more positively to the quality of the record quality than other entities in the same record. The entity reporting worse is then unfairly associated with a high-quality record. ██████████%

[Appendix G](#) presents three tables in which the rankings are based on length ([Table G.1](#)), keywords ([Table G.2](#)) and interesting keywords ([Table G.3](#)). Two horizontal lines can be observed within the rankings. All entities above the top line have reported high quality records on average over a three-year timespan with respect to *Goal 1* in [subsection 7.6.1](#) (removing 90% of the low DQ records). Similarly, entities above the bottom line report high quality records on average, but this is with respect to *Goal 2* in [subsection 7.6.2](#) (finding 90% of the high DQ records). ██████████

██████████ [Figure 7.5](#) shows the counts for the number of records a reporting entity is involved in. ██████████

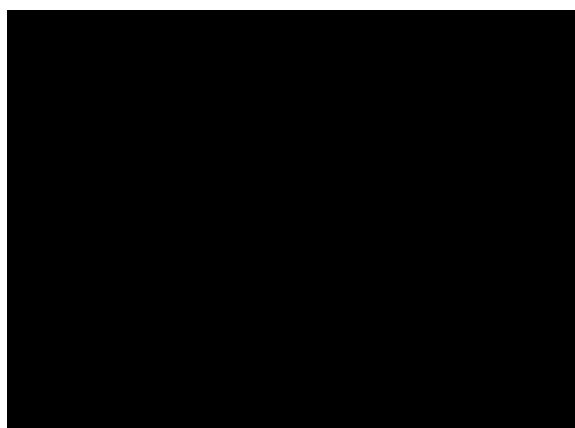


FIGURE 7.5: The counts for the number of records a reporting entity is involved in. Note that this only includes records associated with exactly one reporting entity (72.87% of the records). Records involvements of more than 50 are not visualized, but 15 entities were found. The highest of them has 3037 record involvements.

However, the aforementioned ranking excludes entities that might report extremely

7.11 Timeliness

Timeliness refers to the extent to which the data is sufficiently up-to-date for the task at hand. In order to determine whether records are sufficiently up-to-date, [REDACTED]

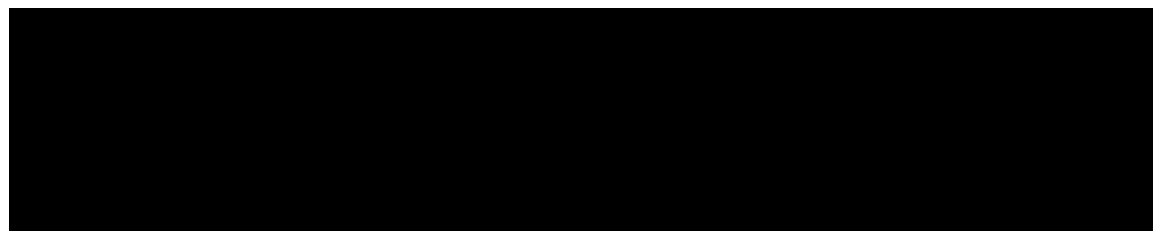
[REDACTED]

Table 7.15 shows some facts about the age distribution of both transactions and records (in days). For records, the ages of its transactions have been averaged. [REDACTED]

[REDACTED]

Table 7.14 shows [REDACTED] $\Delta Days$ columns represent the difference between the reporting date (*Doormelddatum*) and the transaction date (*Transactiedatum*), for both the transaction dates in this research’s dataset and Cognos report⁵ (queried on 05-09-2023). When this value is negative, this means that the transaction date is in the future. The *New rows* column indicates how many new rows were added in Cognos report compared to the old dataset used in this research. Interestingly, for two cases, the records were revised and updated to different dates. However, for one of the two cases, the $\Delta Days$ [REDACTED]

TABLE 7.14: The six records of negative age



[REDACTED] (\approx [REDACTED]).
[REDACTED] (\approx [REDACTED]) [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

⁴ [REDACTED]

⁵Cognos report is a tool to export data from the [REDACTED] police database, in which the most recent version of the STs is available.

[REDACTED]
 [REDACTED]
 [REDACTED] (\approx [REDACTED]) at the moment they are re-
 ported (*Doormelddatum*). [REDACTED] (\approx [REDACTED]). [REDACTED]
 [REDACTED]
 [REDACTED]
 [REDACTED]
 [REDACTED] [Section 1.1](#) mentioned
 that [REDACTED]
 [REDACTED]
 [REDACTED] % [REDACTED]
 [REDACTED] (= [REDACTED]). [REDACTED]
 [REDACTED] % [REDACTED] % [REDACTED]

TABLE 7.15: Facts about the age distribution of the dataset in days

TABLE 8.1: The record list structure containing fictional data

Importantly, the records can be sorted

in [Section 7.6](#).

in [Section 7.6](#).

Moreover, the DQ of records can be associated with reporting entities and groups, which are summarized as well in [Table 8.1](#). As a result, a ranking could be created such as in [Section 7.10](#).

Furthermore, the record list

, further discussed in [Section 8.2](#).

Finally,

8.1.3 User Experience

[REDACTED]

The third advantage is that other investigative bodies and parties in the chain of events (Figure 1.2) utilizing the forwarded records, receive higher quality records. Although it should be investigated to what extent improvements are experienced at other parties, [REDACTED]

[REDACTED]

[REDACTED] subsection 8.1.1 saves the RIK FinEC-ondermijning a tremendous amount of time and frustration.

8.2 Hotspots and Themes

[REDACTED] in Table 8.1 [REDACTED]

[REDACTED] In turn, the municipality might implement administrative approaches to tackle the found insight.

Currently, [REDACTED] in Table 8.1 is [REDACTED]

[REDACTED] Before jumping to conclusions, it is advised to consult Appendix F. [REDACTED]

[REDACTED] % [REDACTED] % [REDACTED]

Chapter 9

Implications and Prospects

The results discussed in [Chapter 7 Results and Discussion](#) have generated new insights that are relevant to the FIU and police. In this chapter, recommendations for both parties are listed. Furthermore, improvement steps for the employed research methodology in [Chapter 6 Methodology](#) are outlined in order to alleviate certain limitations.

9.1 Recommendations

This section presents recommendations targeted at the FIU and police in particular.

9.1.1 FIU

The FIU is the first government instance to receive the transactional data ([Unusual Transactions \(UTs\)](#)) and is thus able to implement improvements on the data quality in an early stage. [Table 9.1](#) introduces an overview of the recommendations applicable to the FIU, along with the associated DQ dimensions responsible for these insights. The recommendations can be stated as follows.

[REDACTED]

[REDACTED] Both the third and fourth recommendation may help to reduce the dataset size.

Fourth, it is recommended to rename certain entity groups, as has been thoroughly discussed in [Section 7.7](#). [REDACTED]

¹Refer to [Section 7.2](#) for more information

Second, theme-based filtering may be an effective method for directing records to specific police districts based on their needs. Within the theme selection, the records with highest DQ should be picked.

[REDACTED]

[REDACTED]

[REDACTED] Statistics may be built using this indicator, such as entity rankings ([Section 7.10](#)).

Fifth, the generated record list in [Section 8.1](#) could be shared with municipalities [REDACTED]

[REDACTED] As a consequence, new administrative approaches, such as a temporary period of performing stricter controls and monitoring, may be formulated for the respective area. However, note that this recommendation is limited by the concern mentioned in [Section 8.2](#). At the moment of writing, the possibilities of realizing recommendations 3-5 in [Table 9.2](#) are explored.

Sixth, it is worth to create additional configuration settings in the dashboard, [REDACTED] or extracting locations of other police units instead of Midden-Nederland. Other RIK FinEC-ondermijning departments in the Netherlands, and other users in general, might benefit from this. Additionally, an option should be included that distinguishes the locations of the reporting entity and reported party as has been proposed in [Section 8.2](#).

Lastly, the labels assigned by the RIK FinEC-ondermijning in [Section 6.3](#) were deemed as the ground truth. However, this begs the question what the ground truth actually is.

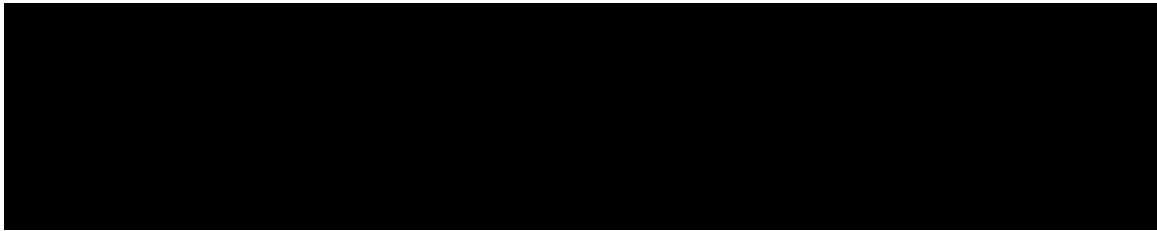
[REDACTED]

9.2 Future Steps

The previous section proposed recommendations based on the results of this research. This section criticizes the research methodology that generated these results and explores potential future steps for improving the current approach and for increasing the depth of [Chapter 7 Results and Discussion](#). The encountered limitations of the employed methodology are acknowledged and new strategies are suggested that may mitigate these constraints. An overview of the limitations and improvement steps is given in [Table 9.3](#).

First of all, the label criteria might have been too vague, introducing proneness to subjectivity. In essence, the criteria were example cases (refer to [Appendix E](#)). On top of that, there was a verbal consensus about the labeling procedure. However, the unstructured nature of this approach might have introduced disparities in labeling behavior among the labelers, caused by differing interpretations of the criteria. This could have been alleviated by including more examples in [Appendix E](#), from which subsequently, certain rules could

TABLE 9.3: Future steps for limitations encountered in this research

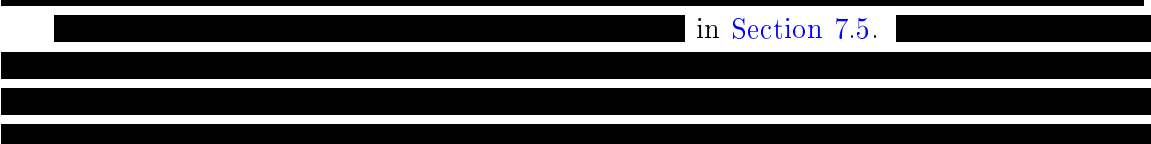


have been extracted and be used for the criteria to strengthen the definitions of the labels.

Second, the coverage of keywords to each theme remains unknown. [REDACTED]

Further research is required to investigate the coverage of the chosen sets of keywords. A strategy might be to label each record with the occurring theme(s) and then to evaluate whether these records are actually covered by the set of keywords. This should illuminate whether some themes are undercovered or overly covered. The latter case seems innocent, but could have affected the performance of the theme-based measures in [Section 7.6](#). Overly covered themes will generally hit on more keywords than undercovered themes and can thus be perceived as an unjustified weight to the theme-based data quality measure. Ideally, the coverage should be equal for all themes by adding keywords for undercovered themes and removing keywords for overly covered themes.

Third, entities report in different ways and thus provide different types of merged record descriptions. [REDACTED]



Sixth, [Section 7.3](#) and [Section 7.5](#) attempted to [REDACTED]




However, this is not guaranteed to be complete (refer to [Appendix F](#)). [REDACTED] [Regular Expression \(RegEx\)](#), [REDACTED]



Seventh, as observed in [Section 7.10](#), [REDACTED]





Lastly, the selected themes and keywords are not future-proof and should be reselected each year since new themes might come up. A more future-proof approach would be to implement machine learning techniques such as topic modeling, which automatically extract themes and keywords from the data in an unsupervised fashion. However, this introduces many other challenges as well (difficulty interpreting themes, requires effective preprocessing steps). Although topic modeling has more potential than manually selecting themes, it requires a higher investment. Nevertheless, it could be more rewarding in the long term.

Chapter 10

Conclusion

In this [Data Quality Assessment \(DQA\)](#), new valuable insights about the [Suspicious Transactions \(STs\)](#) were discovered. The DQA was set out not only to flag data issues, but also to equip interpreters with a deeper understanding of the limitations and potential of the data. This has led to the adoption of a new analytical strategy at the [Regionaal Informatieknooppunt Financieel Economische Criminaliteit \(RIK FinEC\)](#)-ondermijning, allowing them to interpret records more quickly.

Noteworthy data quality findings were found in nearly all of the considered quality dimensions. The granularity of the DQA unveiled data errors that were previously unseen. The lack of entered data has been quantified and ██████████% ██████████

██████████
██████████
██████████
██████████% ██████████
██████████
██████████

The findings of the second research question show that the presence of theme-related keywords in the transaction descriptions reasonably indicate the quality of a record. Besides keywords, the length of a description showed to be an important indicator as well and even outperformed the theme-based methods. A theme-based approach might have more potential when more themes and keywords are included. The coverage of the current set of keywords on the dataset is currently unknown and should be investigated.

In conclusion, recommendations have been proposed to the FIU and police in [Chapter 9 Implications and Prospects](#) as a result of the findings. Additionally, limitations and associated future steps were discussed.

Bibliography

- [1] FD Amicis. A methodology for data quality assessment on financial data. *Studies in Communication Sciences*, 4(2):115–137, 2004.
- [2] De Nederlandsche Bank. Introduction wwft. <https://www.dnb.nl/en/sector-information/supervision-laws-and-regulations/laws-and-eu-regulations/anti-money-laundering-and-anti-terrorist-financing-act/introduction-wwft/>, 2021. (accessed: 17.03.2023).
- [3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):1–52, 2009.
- [4] Kevin Ariza Boekholt. Implementing nlp models & machine learning classifiers to identify high-potential investigations in dutch financial crime. 2023.
- [5] Anti Money Laundering Centre. Jaarrapportage verdachte transacties 2022. 2022. (Confidential).
- [6] Marc de Lignie, Laura Endstra, and Marius Kok. Vinden van financiële onregelmatigheden bij subjecten in politiedata. 2019. (Confidential).
- [7] Martin Domke. Schwifty. github.com/mdomke/schwifty. (accessed: 26.10.2023).
- [8] Martin J Eppler and Peter Muenzenmayer. Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. In *ICIQ*, pages 187–196. Citeseer, 2002.
- [9] Earvin Goudzand. Blinde vlekken. 2019. (Confidential).
- [10] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. Aimq: a methodology for information quality assessment. *Information & management*, 40(2): 133–146, 2002.
- [11] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February 1966.
- [12] J Long, J Richards, and C Seko. The canadian institute for health information (cihi) data quality framework, version 1: a meta-evaluation and future directions. In *Proceedings of the Sixth International Conference on Information Quality*, pages 370–383, 2001.
- [13] Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, mar 2001. ISSN 0360-0300. doi: 10.1145/375360.375365. URL <https://doi.org/10.1145/375360.375365>.

- [14] Customs Administration of the Netherlands. Subversive crime. <https://www.aboutnetherlandscustoms.nl/topics/subversive-crime>. (accessed: 14.04.2023).
- [15] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
- [16] Algemene Rekenkamer. Bestrijden witwassen deel 3: stand van zaken 2021. 2022.
- [17] Friedrich Schneider and Ursula Windischbauer. Money laundering: some facts. *European Journal of Law and Economics*, 26:387–404, 2008.
- [18] Zhanming Su and Zhanming Jin. A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In *Information Quality Management: Theory and Applications*, pages 190–220. IGI Global, 2007.
- [19] FIU the Netherlands. About fiu-the netherlands. <https://www.fiu-nederland.nl/en/home/about-fiu-the-netherlands/>, . (accessed: 04.04.2023).
- [20] FIU the Netherlands. What is money laundering? <https://www.fiu-nederland.nl/en/about-the-fiu/what-is-money-laundering>, . (accessed: 09.03.2023).
- [21] FIU the Netherlands. Reporting groups. <https://www.fiu-nederland.nl/en/meldergroepen>, . (accessed: 17.03.2023).
- [22] FIU the Netherlands. What is terrorism financing? <https://www.fiu-nederland.nl/en/home/about-fiu-the-netherlands/what-is-terrorism-financing/>, . (accessed: 05.04.2023).
- [23] FIU the Netherlands. Wwft (prevention) act. <https://www.fiu-nederland.nl/en/legislation/general-legislation/wwft>, . (accessed: 17.03.2023).
- [24] FIU the Netherlands. Annual reports. 2006-2020.
- [25] FIU the Netherlands. Annual review of fiu-the netherlands. <https://www.fiu-nederland.nl/wp-content/uploads/2023/02/FIU-Annual-Review-2021.pdf>, 2021. (accessed: 08.11.2023).
- [26] FIU the Netherlands. Annual review of fiu-the netherlands. <https://www.fiu-nederland.nl/wp-content/uploads/2023/07/FIU-Annual-review-2022-ENG-web.pdf>, 2022. (accessed: 08.11.2023).
- [27] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. Zicht op ondermijning. <https://www.zichtopondermijning.nl>. (accessed: 26.07.2023).
- [28] Yair Wand and Richard Y Wang. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95, 1996.
- [29] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.
- [30] P Zhang and G Chartrand. *Introduction to graph theory*. Tata McGraw-Hill, 2006.

Appendix A

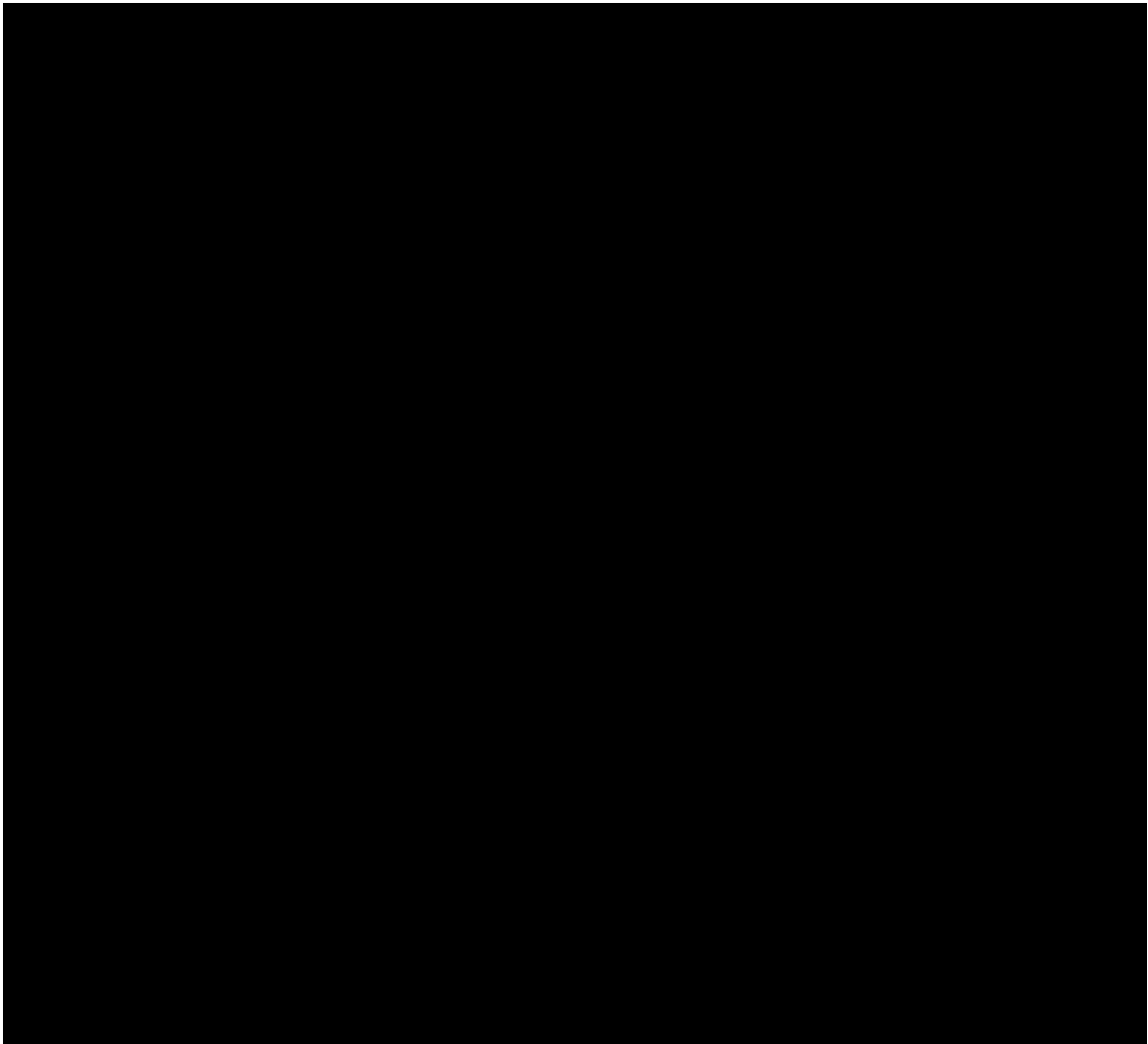
Reporting Entities

1. Accountants
2. An entity that is not a bank, but carries out banking activities
3. Art dealers
4. Banks
5. Casinos
6. Civil-law notaries
7. Dealers in goods
8. Dealers or brokers in high-value goods
9. Electronic money entities
10. Investment entities
11. Investment firms
12. Lawyers
13. Legal service providers
14. Life insurance
15. Life insurance brokers
16. Money exchange entities
17. Natural or legal persons that put their address at another's disposal (domicile-providers)
18. Pawn shops
19. Payment service broker
20. Payment service provider
21. Professional or commercial providers of custodian wallets
22. Professional or commercial providers of services for the exchange between virtual currencies and fiduciary currencies
23. Providers of remote gaming services
24. Real estate agents
25. Safe custody services
26. Tax advisors
27. Trust offices
28. Undertaking for Collective Investment in Transferable Securities
29. Valuers

Appendix B

Dataset Structure

TABLE B.1: The dataset structure



Appendix C

Dimension Matrix

- d_1 = Completeness
- d_2 = Concise representation
- d_3 = Consistent representation
- d_4 = Ease of manipulation
- d_5 = Free-of-error
- d_6 = Value-added
- d_7 = Understandability
- d_8 = Objectivity
- d_9 = Believability
- d_{10} = Reputation
- d_{11} = Timeliness

TABLE C.1: An overview of how the dimensions apply to specific columns. A cell marked with an x means that the column can be used to compute the dimension score. For completeness, 100% means that it is always complete. For free-of-error, 100% means it is always correct.

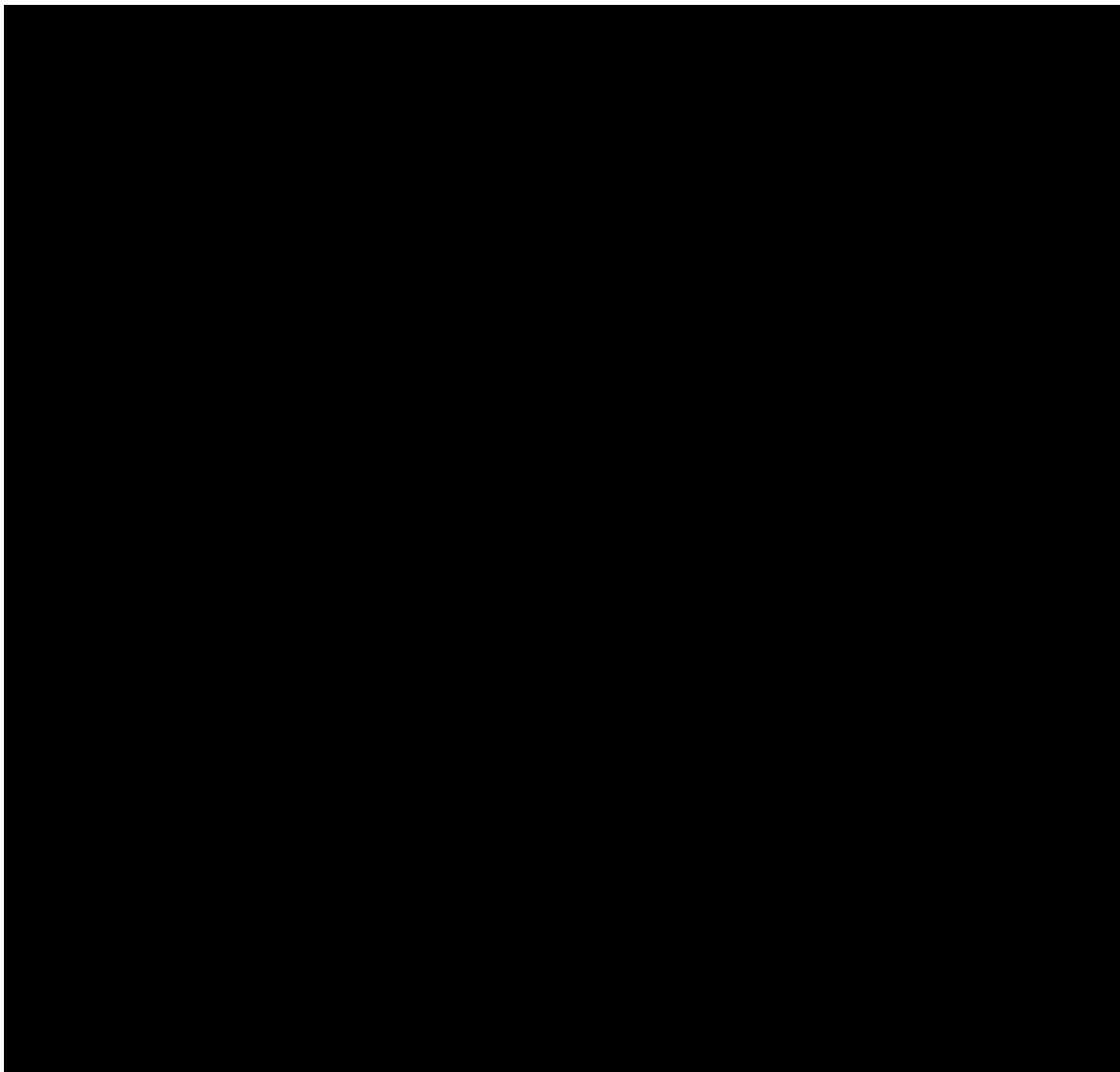
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
██████████	100%			x	100%						
██████████████	100%			x	100%				x		
██████████████	x			x	100%				x		
██████████████████	100%			x	100%						
██████████████	100%			x	100%						x
██████████████	x		x	x				x	x		
██████████████	100%			x	100%				x		x
██████████████	100%	x		x	100%				x		
██████████████████	x		x	x					x		
██████████	x		x	x	x				x		
██████	x		x	x					x		
██████	100%			x	100%				x		
██████████	x			x	100%				x		
██████████████	x			x	100%				x		
██████████████████	x	x		x		x		x	x	x	
██████████████	100%			x	100%				x	x	
██████████████	x			x	100%		x		x	x	
██████████████	100%			x	100%		x		x		

████████████████████	100%			x	100%		x		x		
██████████	100%			x	100%						
██████████	x	x		x					x		
██████████	x	x		x					x		
██████████████	x			x	x				x		
████████████████	x		x	x					x		
██████████	x		x	x	x				x		
██████████	x		x	x					x		
██████████	x			x	100%				x		
██████	x		x	x					x		
██████████████	x		x	x	x				x		
██████████████	x			x					x		
██████████████	x			x					x		
████████████████████	x		x	x	x				x		
██████████████	100%			x	100%						
██████████████	x			x			x				

Appendix D

Search Terms

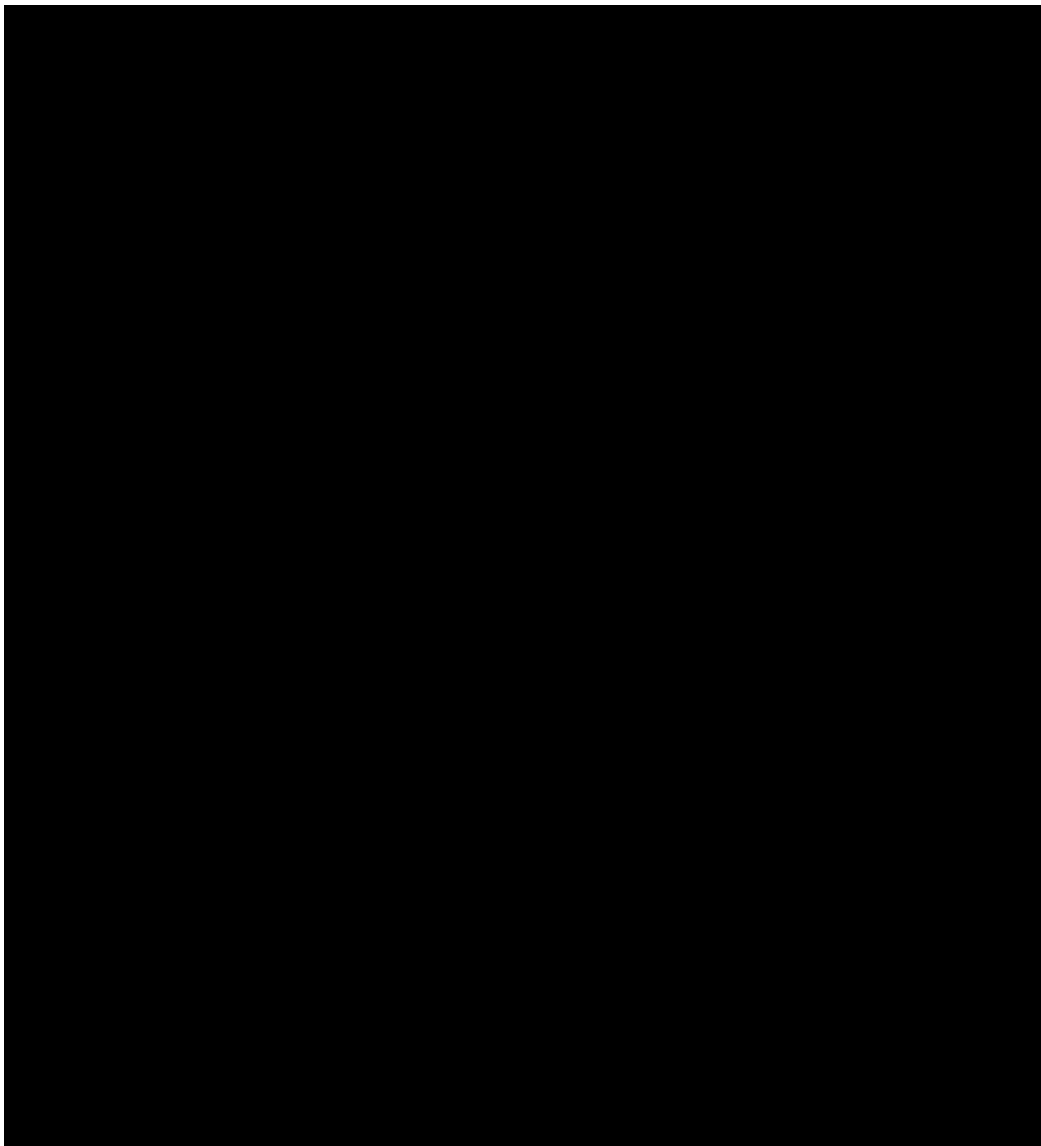
TABLE D.1: Themes/phenomena and their corresponding search terms. When *AMLC* is set to *No*, it is a new theme in addition to the AMLC [5] themes. Bold-fied keywords are new as well. *Interesting* indicates whether the RIK FinEC-ondermijning finds it interesting or not.



Appendix E

Labeling Criteria

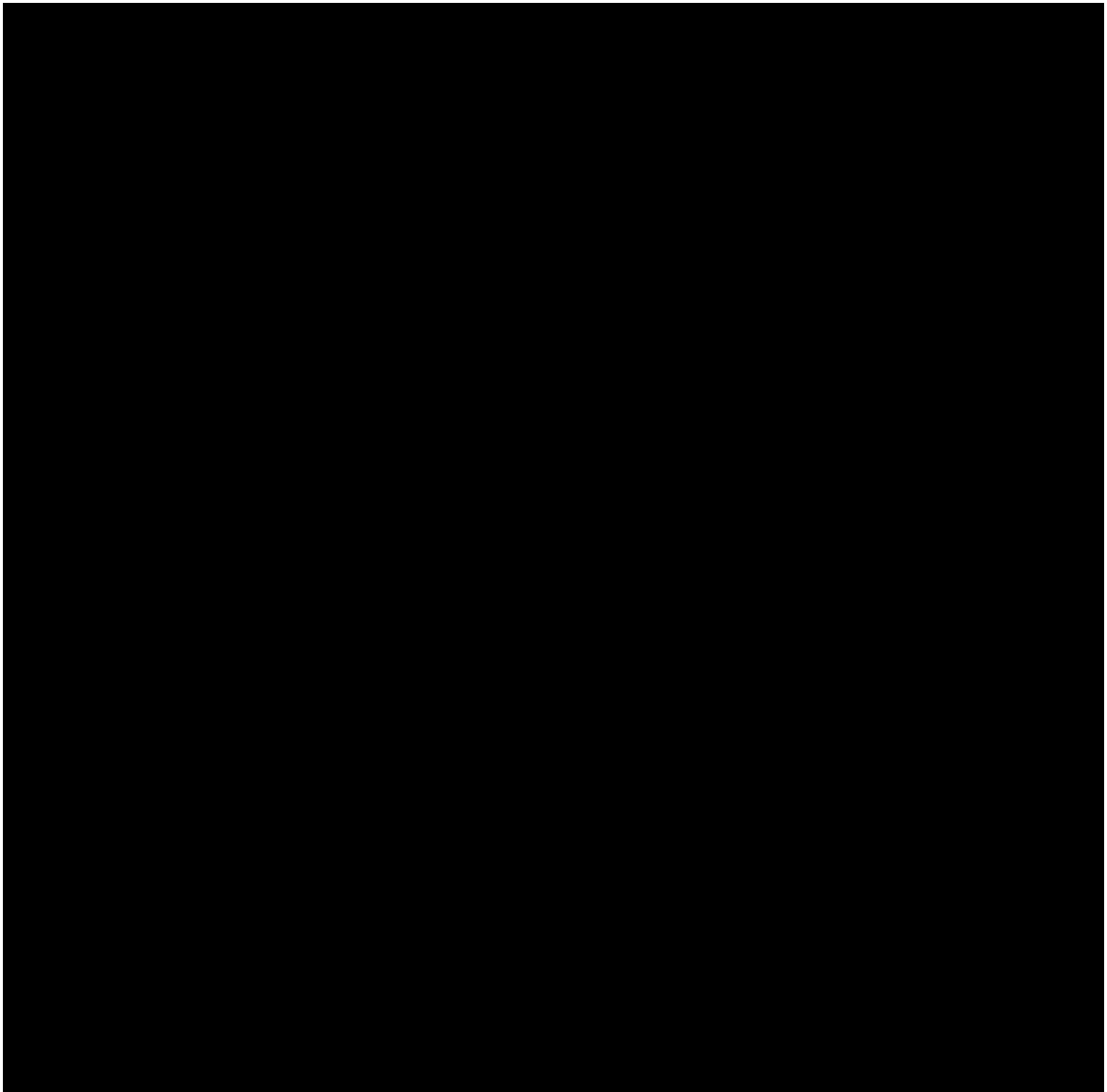
TABLE E.1: Labeling criteria



Appendix F

Completeness

TABLE F.1: Completeness results



Appendix G

Reputation

TABLE G.1: The reporting entities (involved in at least 10 records) ranked based on the average merged description length over a three-year timespan. The top horizontal line is the *Goal 1* threshold in [subsection 7.6.1](#). The bottom line is the *Goal 2* threshold in [subsection 7.6.2](#).

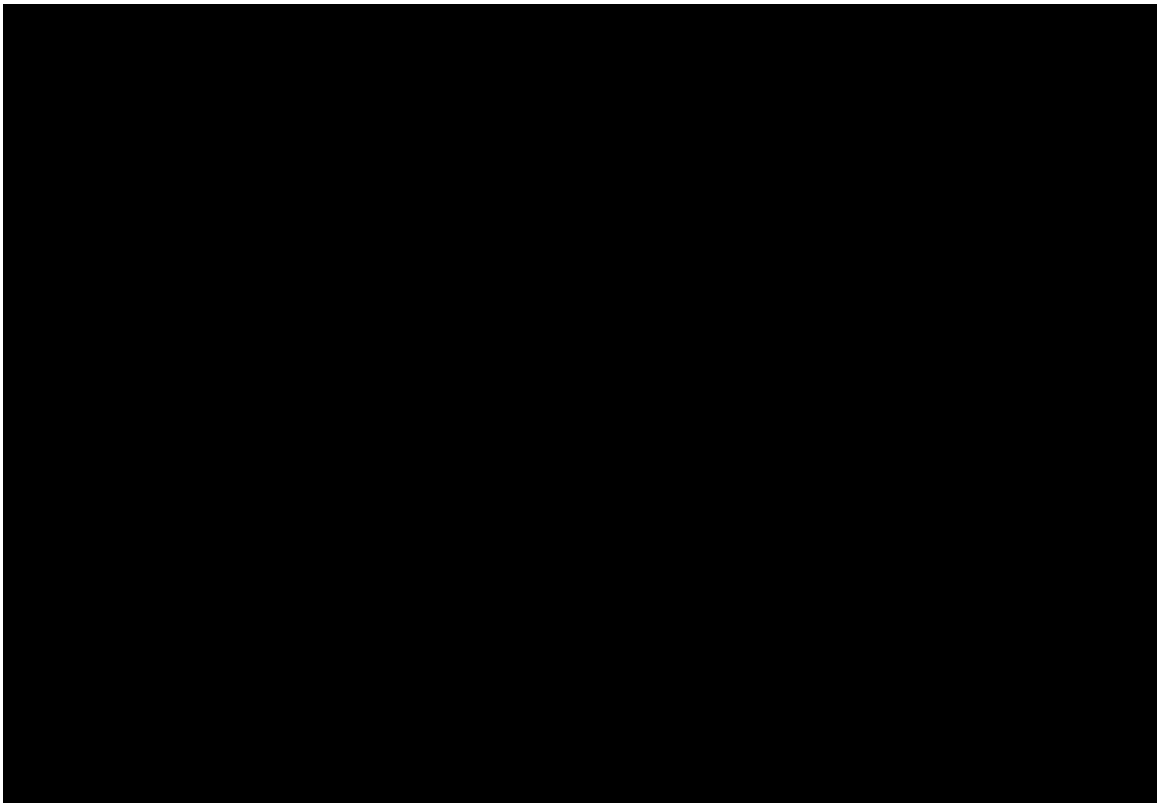


TABLE G.2: The reporting entities (involved in at least 10 records) ranked based on the average number of keywords in the merged description over a three-year timespan.

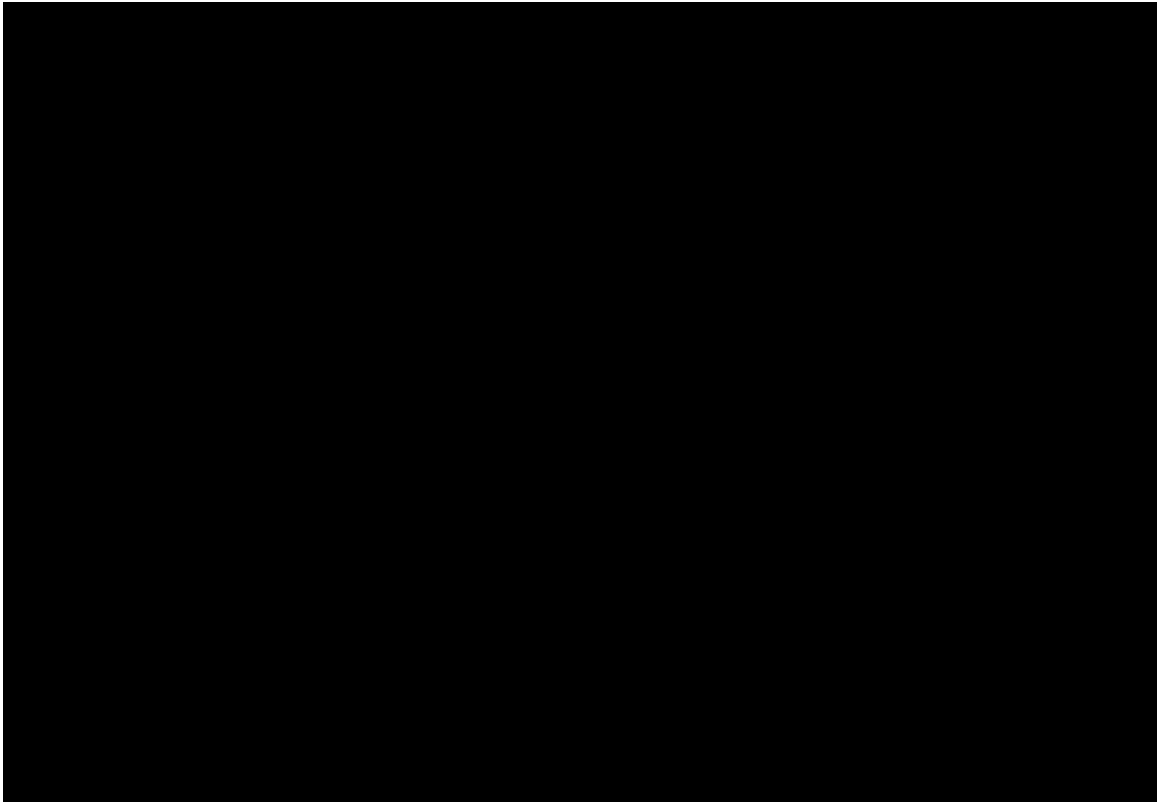
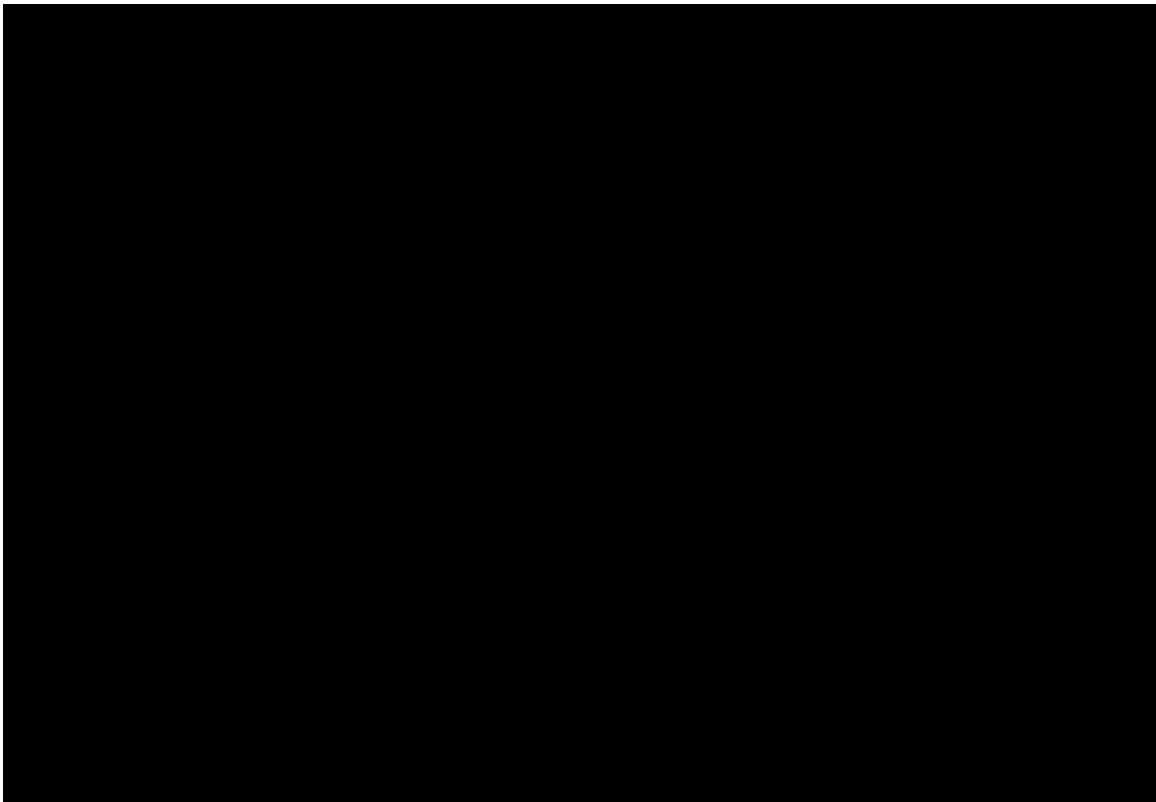


TABLE G.3: The reporting entities (involved in at least 10 records) ranked based on the average number of interesting keywords in the merged description over a three-year timespan.



Appendix H

Threshold Evaluation

TABLE H.1: Performance of evaluation metrics with respect to target scores




TABLE H.2: Performance of evaluation metrics with respect to target scores

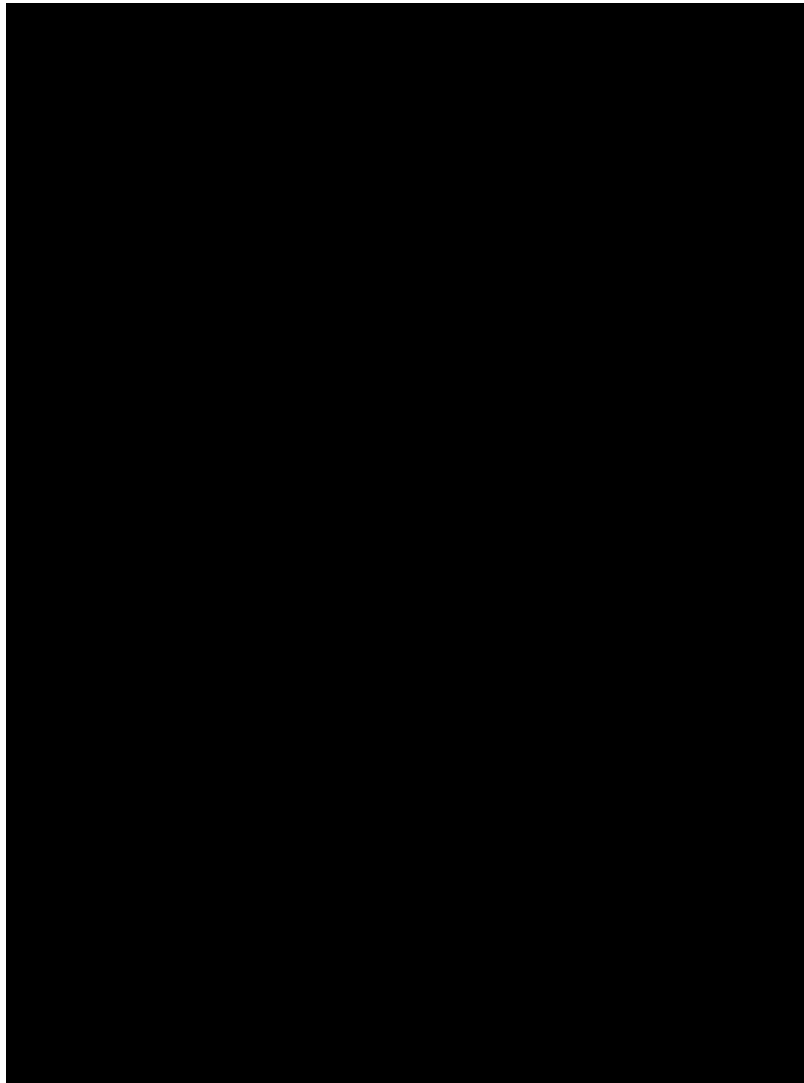



TABLE H.3: Performance of evaluation metrics with respect to target scores



TABLE H.4: Performance of evaluation metrics with respect to target scores



Appendix I

Effectiveness of Social Network Analysis

This section contains a small research cycle about the effectiveness of [Social Network Analysis \(SNA\)](#) to the [STs](#) dataset. A summary of the research results is provided in [Section I.4](#).

I.1 Goal and Approach

Recall that records can be viewed as social networks in which transactions define relationships between parties. SNA algorithms, such as centrality measures, can then be applied to reveal patterns and insights. Centrality measures indicate the most important, central or influential node in the network, which is useful for finding the key player.

The goal of this small research cycle is to explore to what extent centrality measures are effective with respect to this dataset. [REDACTED]

I.2 STs as Graphs

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a mathematical structure composed of a set of vertices \mathcal{V} interconnected by a set of edges \mathcal{E} . A graph can be used to represent a list of objects that have connections between them [30]. When a graph consists of relations between persons, groups or organizations, it is called a social network. Note that the terms vertex and node will be used interchangeably.

All of the STs could be regarded as one big graph in which a vertex v represents a person or bank account. An edge e on the other hand could represent the transaction itself between two vertices. A transaction typically includes a sender and receiver. This would mean that in a transaction network each edge has a direction. Such an edge is called a *directed edge*. All of the edges are directed, so the graph itself is called a *directed graph*. Each of the edges can be assigned a value, which is the amount of money involved in the respective transaction. The graph is then called a *weighted graph*. [REDACTED]

[REDACTED] G_{sub} [REDACTED]
[REDACTED]

I.3 Results and Discussion

After exploring different records from the STs dataset, four categories of challenges were identified. Note that references have been provided to specific records. These are formatted as: [REDACTED]

I.3.1 Small Subgraphs

This subsection will evaluate whether there is a potential risk of small subgraphs. The smaller a subgraph is, the less meaningful SNA is. For example, finding the key player in a network of only two nodes does not provide a significant insight. Each record can be regarded as a subgraph of a large [REDACTED] graph (representing the total dataset). These subgraphs vary in size and structure. [REDACTED]

Investigating whether the risk of small subgraphs is real can be achieved through various approaches. The first one would be [REDACTED]

[REDACTED]

Another approach is to look at [REDACTED]

[REDACTED]

Although the previous approaches have some inaccuracies, they will still be used to obtain an approximation of the graph sizes. [Table I.1](#) [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED] This would result in an incorrect graph ([Figure I.1](#)) [REDACTED]

[REDACTED]

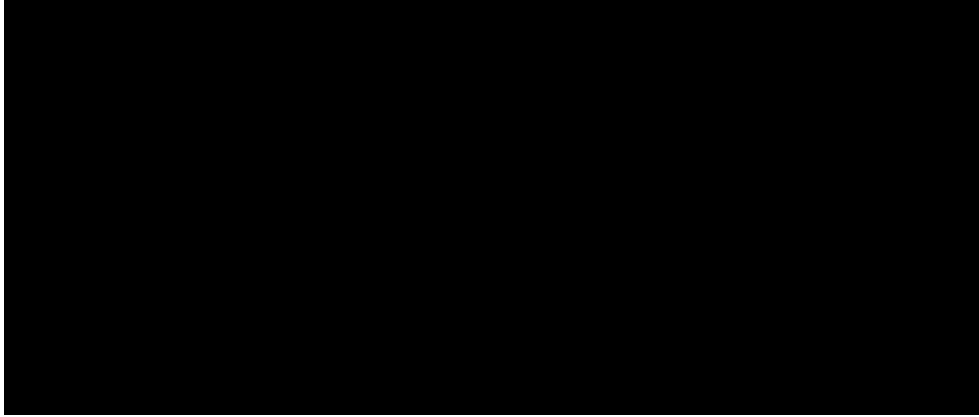
[REDACTED]

FIGURE I.1: The receiver is not specified.

On the other hand, the rest of the means in [Table I.1](#) [REDACTED]

[REDACTED] (explained later on in [subsection I.3.3](#) and [subsection I.3.4](#)). [REDACTED]

TABLE I.1: Summary statistics of the dataset viewed per record.



I.3.2 Ego Network

[REDACTED]

[REDACTED] A randomly sampled record from the dataset is depicted in [Figure I.2](#) [REDACTED]

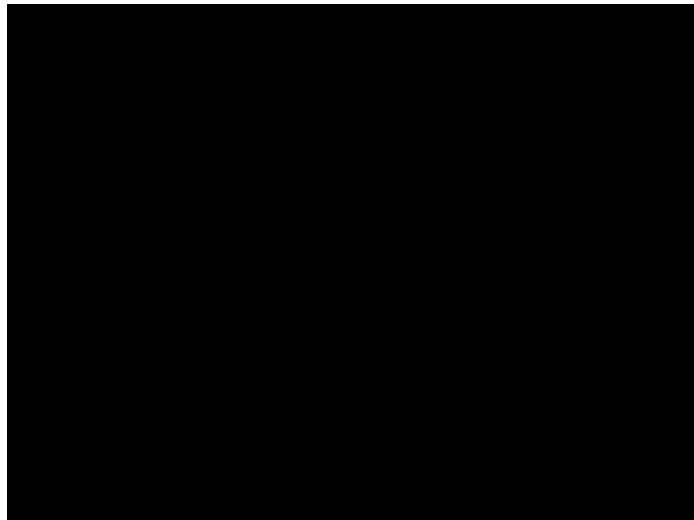


FIGURE I.2: The ego is obviously v_1 .

I.3.3 Groups

[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED] Therefore, the results in [Table I.1](#) should be carefully interpreted.

I.3.4 Semantic Duplicates

Semantic duplicates are data entries that have the same meaning, while being not exactly equal to each other. [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

[REDACTED] For example, [REDACTED], but [REDACTED]. To link back to [Table I.1](#), the party mean is [REDACTED]
[REDACTED]

I.4 Summary

To summarize, the goal of this research cycle was to identify the challenges of [SNA](#) when applied to [STs](#). Records can be interpreted as social networks, or rather weighted directed graphs, [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]
[REDACTED]

List of Abbreviations

- A&O** Analyse en Onderzoek. [5](#)
- AAP** Advanced Analytics Platform. [47](#)
- AMLC** Anti Money Laundering Centre. [18](#), [24](#)
- DQ** Data Quality. [4](#), [9](#), [19](#), [27](#)
- DQA** Data Quality Assessment. [4](#), [8](#), [11](#), [12](#), [19](#), [27](#), [55](#)
- FIU** Financial Intelligence Unit. [1](#), [2](#), [13](#), [16](#)
- IBAN** International Bank Account Number. [21](#), [30](#)
- KenO** Kenteken- en Opsporingsysteem. [28](#), [50](#)
- KvK** Kamer van Koophandel. [21](#)
- RegEx** Regular Expression. [21](#), [53](#)
- RIK FinEC** Regionaal Informatieknooppunt Financieel Economische Criminaliteit. [2](#), [3](#), [4](#), [5](#), [6](#), [55](#)
- SNA** Social Network Analysis. [72](#), [75](#)
- STs** Suspicious Transactions. [1](#), [10](#), [16](#), [19](#), [55](#), [72](#), [75](#)
- UTs** Unusual Transactions. [1](#), [50](#)
- Wwft** Wet ter voorkoming van witwassen en financieren van terrorisme. [2](#), [6](#)