Supporting Decision-making in Fraud Sensitive Environments

Including Personal Data from Public Sources in Risk Analyses

Yorick Bouma

Universiteit Utrecht

Department of Information and Computing Sciences

Utrecht University

Competence Center Business Intelligence

Info Support

July 16, 2014

ABSTRACT

The purpose of this research is to investigate if, and how, personal data from public sources can be utilized for risk analyses of (prospective) customers. An objective was to identify the steps necessary for the implementation of a system that enables this, and what the architecture of such a system would be. An additional objective was to find out what legal issues arise with the implementation of such a system in the Netherlands.

For this investigation, a qualitative approach was taken. A literature study on Business Intelligence, Web Information Extraction and Entity Matching was performed in order to identify techniques and methods with which the objectives could be attained. Additionally, to acquire knowledge about the current state of risk analyses, three interviews with experts in the field were conducted. Based on the data that was gathered during these activities, the Public Sources for Risk Analyses (PSRA) Process and PSRA Architecture were constructed.

The PSRA Process contains six phases, each containing several steps. The purpose of this process is to guide the development of a system that extends current risk analysis systems with personal data from public sources. The PSRA Architecture, in turn, serves as a high-level reference architecture for such a system.

These two artifacts were evaluated by the implementation of a proof of concept, and by experts in the field. The proof of concept was unable to prove that personal data from public sources can be utilized for risk analyses, since it could not accurately decide which profile belonged to a particular subject. This was mainly caused by the lack of publicly available personal data. Additionally, some legal obligations that should be met in the Netherlands make the utilisation of personal data from public sources in risk analysis systems even more difficult.

DEDICATION

Lovingly dedicated to my mother, father, girlfriend and friends.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisors from the Department of Information and Computing Science at Utrecht University. My first supervisor — dr. Marco Spruit — for his guidance, advice in decisions, and constructive feedback during the research project as well as during the writing of my thesis. And my second supervisor — prof. dr. Sjaak Brinkkemper — for his constructive feedback during the writing of my thesis.

Second, I would like to thank all my supervisors from Info Support. My main supervisor – Hans Geurtsen – and technical supervisor – Koos van Strien – for their guidance, advice in descisions, and constructive feedback during the research project as well as during the writing of my thesis. And my process supervisor – Pascalle Hijl – for her guidance during the research project.

Last but not least, I would like to acknowledge the support provided by my family and friends during the research project and the writing of my thesis.

CONTENTS

I OPENING 9

1 INTRODUCTION 10
- 1.1 Problem Statement 10
- 1.2 Research Question 11
- 1.3 Relevance 11
- 1.4 Thesis outline 12

2 METHODOLOGY 13
- 2.1 Design Science Research 13
- 2.2 Scope 15
- 2.3 Research Approach 16

II RESULTS 22

3 LITERATURE REVIEW 23
- 3.1 Business Intelligence 23
- 3.2 Web Information Extraction 29
- 3.3 Entity Matching 32
- 3.4 Related Studies 36
- 3.5 Legal issues 39

4 EXPLORATORY INTERVIEW RESULTS 44
- 4.1 Fraud and the company 44
- 4.2 Manual measures 45
- 4.3 Automated systems 46
- 4.4 Manual use of personal data from public sources 47
- 4.5 Automated use of personal data from public sources 48
- 4.6 Ethical issues 50

5 PUBLIC SOURCES FOR RISK ANALYSES 52
- 5.1 PSRA Process 52
- 5.2 Architecture 75

III EVALUATION 82

6 EVALUATION THROUGH A PROOF OF CONCEPT 83
- 6.1 Process 83
- 6.2 Architecture 100
- 6.3 Limitations 106
- 6.4 Future Extensions 108

7 IMPROVEMENTS 110

IV CLOSURE 112

8 CONCLUSIONS 113

9 DISCUSSION 115

10 FUTURE RESEARCH 116

11 REFERENCES 118

V APPENDICES 121

A PSRA PROCESS 122

B INTERVIEW PROTOCOL 123
- B.1 Introduction 123

B.2 Fraud 123
B.3 Fraud detection and prevention 123
B.4 Use of personal data from public sources 123
B.5 Ethical issues 124
C INTERVIEW SUMMARIES 125
D DECISION FUNCTION WEIGHTS 133

LIST OF FIGURES

Figure 2.1 Information Systems Research Framework 14
Figure 2.2 Business Intelligence Framework 15
Figure 2.3 The three stages of effective literature process 17
Figure 3.1 Business Intelligence architecture adopted from Chaudhuri and Dayal (1997) 24
Figure 3.2 Business Intelligence architecture adopted from Negash (2004) 25
Figure 3.3 Business Intelligence architecture adopted from Watson and Wixom (2007) 25
Figure 3.4 Business Intelligence architecture adopted from Chaudhuri, Dayal, and Narasayya (2011) 25
Figure 3.5 Business Intelligence architecture adopted from Turban, Sharda, Delen, and King (2012) 26
Figure 3.6 Business Intelligence architecture adopted from Info Support 26
Figure 3.7 High-level architecture of the "getting data in" part of a Business Intelligence system 27
Figure 3.8 Extract, transform, load process for personal data from web sources 28
Figure 3.9 Web Information Extraction architecture 30
Figure 3.10 Wrapper architecture 32
Figure 3.11 Entity matching for profiles on public sources 36
Figure 5.1 Phases of the PSRA Process 52
Figure 5.2 The Legal Understanding phase 54
Figure 5.3 The Attribute Identification phase 57
Figure 5.4 The Source Selection phase 61
Figure 5.5 The Web Information Extraction phase 64
Figure 5.6 The Entity Matching phase 68
Figure 5.7 The System Construction phase 72
Figure 5.8 Functional reference architecture of a PSRA System 76
Figure 5.9 Technical reference architecture of a PSRA System 76
Figure 5.10 Functional reference architecture of the wrapping functionality 77
Figure 5.11 Technical reference architecture of the wrapper module 78
Figure 5.12 Functional reference architecture of the matching functionality 80
Figure 5.13 Reference architecture of the matcher module 80
Figure 6.1 Architecture of the proof of concept 101
Figure 6.2 Architecture of the wrapper module of the proof of concept 102
Figure 6.3 Data Vault data model for Facebook 104
Figure 6.4 Architecture of the matcher module of the proof of concept 106

Figure A.1 The PSRA Process 122

LIST OF TABLES

Table 2.1 Experts 20
Table 2.2 Expert feedback experts 21
Table 3.1 Comparison of related studies and this research 37
Table 3.2 Relevant indexes on `www.codices.coe.int` 40
Table 5.1 Example deliverable of the Attribute Identification phase 60
Table 5.2 Example list of locally available public sources in the Netherlands (not a complete list) 61
Table 5.3 Example matrix of the Calculate attribute fulfillment step 62
Table 5.4 Example prioritized list of locally available public sources in the Netherlands (not a complete list) 63
Table 5.5 Confusion matrix 74
Table 6.1 Attributes extracted from the interview summaries 88
Table 6.2 Prioritized list of attributes 88
Table 6.3 List of locally available public sources identified in the Netherlands 89
Table 6.4 Matrix with attribute fulfillment scores 90
Table 6.5 Prioritized list of locally available public sources in the Netherlands 90
Table 6.6 Weights for the LinkedIn weighted average decision function 97
Table 6.7 Confusion matrix for Facebook 99
Table 6.8 Confusion matrix values for LinkedIn, Google+ and Twitter 100
Table C.1 Attributes identified in Interview I 126
Table C.2 Attributes identified in Interview II 128
Table C.3 Attributes identified in Interview III 130
Table D.1 Weights for the LinkedIn weighted average decision function 133
Table D.2 Weights for the Facebook weighted average decision function 133
Table D.3 Weights for the Twitter weighted average decision function 133
Table D.4 Weights for the Google+ weighted average decision function 134

Part I

OPENING

1

INTRODUCTION

Making the right decisions within a company is obviously important, also in fraud sensitive environments. One wrong decision can have huge financial consequences, let alone low quality decision-making that is automatically repeated many times in, for example, an operational system. According to Turban et al. (2012) these "decisions may require considerable amounts of relevant data, information and knowledge", which will form the foundation of the decisions. Gathering this relevant data, information and knowledge may be difficult, certainly when the decision-making concerns a prospective customer of the business (B2C). After all, there will be practically no information available within the company about this customer, and accessing other companies internal information might be difficult.

At the same time, the amount of publicly available information on the Internet is "huge and still grows" (Liu, 2007). Each day, more information is added, including information about these prospective customers. The Internet is filled with information and contains traces that people leave behind during their on-line activities. It would be a major benefit when this publicly available information can be used as a foundation to support those decisions.

1.1 PROBLEM STATEMENT

In an ideal situation, all information related to a prospective customer of the business is available during the decision making process in a fraud sensitive environment. This way, a fully informed decision can be made, with relevant information to substantiate the actual decision. Risk analyses of prospective customers are used to support this decision-making.

Currently, these risk analyses are already done with the use of information from internal and private sources. Unfortunately, information about the prospective customer may not always be available in these internal and private sources. And if there is some information available, this does not automatically mean that it is always sufficient to support the decision making process. Because of this incomplete or missing information, wrong decision can be made.

By including the information that is available on public sources in these risk analyses, both of these issues are addressed. In the case that no internal information exists at all, publicly available information gives decision-makers at least some information to work with. In the case that there is already internal information available for the risk analysis, the publicly available information can be used as an extension and as validation. Therefore, this research tries to narrow the gap between the ideal and the current situation with a system that extends a current risk analyses system with personal data from public sources. A process and an architecture that, respectively, guide

the implementation of and serve as a reference architecture for such a system.

1.2 RESEARCH QUESTION

From the problem statement described in the previous section, the main research question of this research is formulated as follows:

> How can personal data from public sources be utilized for risk analyses of (prospective) customers and thereby support decision-making in fraud sensitive environments?

In order to answer the main research question, and in order to build the process and architecture, the following sub-questions are formulated:

SQ1 Which steps should be taken in order to include personal data from public sources in a risk analysis system?

SQ2 What would be the architecture of a system that includes personal data from public sources in a risk analysis system?

Additionally, the third sub research question originates from the business need to identify legal issues, and is formulated as follows

SQ3 What are the legal issues that arise when personal data from public sources is used in the Netherlands?

1.3 RELEVANCE

SCIENTIFIC RELEVANCE First, little to none scientific papers exist about how publicly available information can be utilized for risk analyses. This study will attempt to identify and describe the possibilities and propose a process that guides the implementation such a system that makes use of public sources.

In addition, very little is written in scientific papers about legal issues that arise when one utilizes personal data from public sources to support decision-making. This is a subject that is certainly an issue at the moment, since more and more information is collected and combined to do analysis on. This study will identify the legal issues within the Netherlands, and a part of the proposed process focuses on how legal issues in other implementation locations can be identified.

Third, two artifacts in the form of a process and a reference architecture will be created during this research, thereby contributing directly to the scientific knowledge base. These artifacts can be used for future research, for instance, to improve or extend the process. It will also identify new gaps that can be addressed by future research.

BUSINESS RELEVANCE The relevance of this research can also be seen from a business perspective. Businesses are already doing risk analyses based on personal data from internal and private sources, personal data from public sources could enhance

the overall risk analyses. Since these risk analyses support the decision-making process, this will also be enhanced.

Improving the decision making process, even marginally, can be of great value to a company. It could potentially save a lot of money, because the decision making process is often part of the daily process in companies. Every wrong decision that is prevented by being more informed, with the use of personal data from public sources data, can potentially save money.

1.4 THESIS OUTLINE

The remainder of this thesis is structured as follows. The subsequent chapter — the last chapter of the opening part — will discuss the methodology that was used to conduct this research and the approach that was taken.

The second part presents the results of this study in three chapters. The first chapter presents the results from the conducted literature study. Hereafter, the result from the exploratory interviews are shown. Finally, the constructed artifacts - the PSRA Process and PSRA Architecture - are presented.

The Evaluation part, the second to last part, contains two chapters. The first chapter in this part describes the implementation of the proof of concept. This implementation evaluates the process as well as the reference architecture. Hereafter, a chapter is dedicated to discuss the points for improvements that were identified by experts in the field and by the analysis of the prototype.

The last part of this thesis, the closure, presents the conclusion of this study. Additionally, it addresses some liabilities of the research in a discussion chapter and directions for further work in a future research chapter.

2

METHODOLOGY

2.1 DESIGN SCIENCE RESEARCH

Most research in the Information System Discipline is characterized by two paradigms: behavioral science and design science (Hevner, March, Park, & Ram, 2004). The first paradigm, behavioral science, aims to develop and verify theories. Theories focused on explaining or predicting human or organizational behavior. The design science paradigm, the one applicable to this research, aims to create new and innovative artifacts. Artifacts that are focused on extending the boundaries of human and organizational capabilities. Artifacts are defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems) (Hevner et al., 2004). The design science paradigm is applicable to this research since two new artifacts will be created, namely the PSRA Process and the PSRA Architecture.

Because the design science paradigms matches this research, the Information Systems Research Framework (depicted in Figure 2.1) and the associated seven guidelines are used (applied in Section 2.1.1). The research consists of constructing the PSRA Process and PSRA Architecture (artifacts). In addition, this process will be executed and thereby a proof of concept will also be build. The architecture will be used as a reference architecture for this proof of concept. One iteration of the assess and refine cycle will be executed. The proof of concept will be analyzed to identify points for improvements and experts in the field will also evaluate the process and architecture. Based upon these identified points for improvement the process and architecture will be refined once before finalizing the research project.

The knowledge base consists of theories, frameworks, instruments and artifacts from the four literature study topics: business intelligence, web information extraction, entity matching and legal issues. These will be used as a foundation for the information science research. In order to conduct a profound research, several methodologies from the knowledge base will be utilized. The research project will contribute to the knowledge base by adding two artifacts, namely the proposed process and reference architecture.

The environment consists of people and organizations that are active in fraud sensitive environments. From these people and organizations certain business needs influence the research project. Initially, the demand for this research originates from the business. The study of legal implications that arise when implementing a system that utilizes personal data from public sources within the Netherlands, was part of the business needs. The technology in the environment to which the research will be applicable are risk analysis systems. These could be improved by the findings of this study by including publicly available personal data.

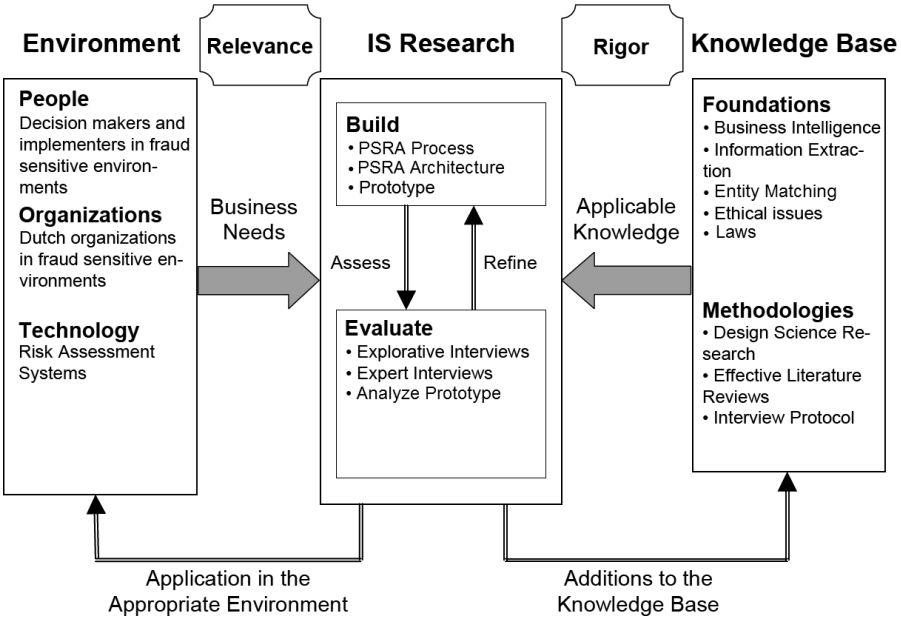

Figure 2.1.: Information Systems Research Framework (Hevner et al., 2004) applied to this research

2.1.1 *Guidelines*

DESIGN AS AN ARTIFACT The research produces two viable artifacts in the form of the PSRA Process and the PSRA Architecture. The process guides the implementation of a system that extends existing risk analysis systems with personal data from public sources. The architecture will serve as a reference architecture for such a system.

PROBLEM RELEVANCE The objective of this research is to develop a solution to the problem of having few or no information in a risk analysis for proactive consumers. Additionally, it also aims to increase the quality of a general risk analyses by including publicly available information.

DESIGN EVALUATION The artifacts, the PSRA Process and PSRA architecture, will both be evaluated. A proof of concept will be implemented by executing the process. The PSRA Architecture will be used as the reference architecture for the architecture of that proof of concept. Additionally, experts in the field will also evaluate the process and reference architecture. From this evaluation, points for improvement will be identified for both the process and the architecture.

RESEARCH CONTRIBUTIONS The contributions of the research are the process and reference architecture. Additionally, it will also contribute on a theoretical level by including legal issues associated with the implementation of a system that utilizes personal data from public sources,. They extend the knowledge base and are built upon existing knowledge.

RESEARCH RIGOR The existing knowledge base is used to extract theoretical foundations and research methodologies. These are then used to build upon or used as guidelines in the research.

DESIGN AS A SEARCH PROCESS The PSRA Process and PSRA Architecture will be built based on the artifacts identified in the literature study and exploratory interviews. By the implementation of a proof of concept and the evaluation by experts in the field an additional cycle of assessing and refining is completed.

COMMUNICATION OF RESEARCH The research will be presented effectively to both technical-oriented and management-oriented audiences. Although the implementation of an information system is a technically matter normally, the proposed process and reference architecture are mostly high-level and thus understandable for a less technical audience as well. When necessary, more detailed and technical information is supplied in this thesis.

2.2 SCOPE

To ensure a thorough research project the scope will be limited in comparison to the entire process of risk analysis based on public available personal data. The focus will be on the first part of the process, wherein personal data from public sources is extracted, linked to the subject of the risk analysis and loaded into the data storage of the existing risk analysis system. In order for this to be possible, the existing risk analysis system should contain a data storage, for example a data warehouse, wherein the data can be loaded. Therefore, it is assumed that this existing risk analysis system is a business intelligence system with an associated data warehouse.

The business intelligence framework of Watson and Wixom (2007), depicted in Figure 2.2, makes this dichotomy clear. Watson and Wixom (2007) describe that "BI is a process that includes two primary activities: getting data in and getting data out". In this context the left side, getting data in, will be the primary focus of the research project. The right side, getting data out, will be left for future research.

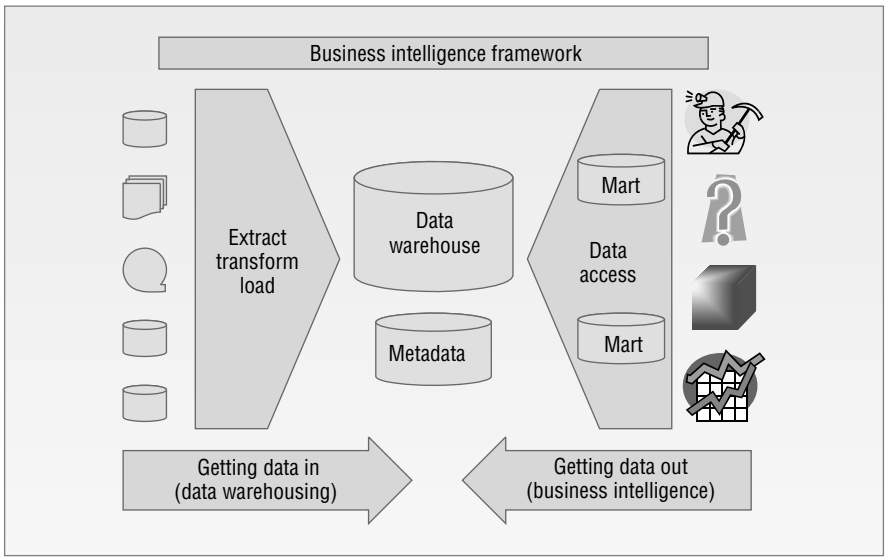

Figure 2.2.: Business Intelligence Framework (Watson & Wixom, 2007)

The choice of focus on the first part occurs from several reasons. First, it seems logical to start with the part of the process that comes first. This builds a solid base for research on the second part. Doing it the other way around seems illogical, since it should then be assumed that the first part is already researched while this is not the case.

Secondly, the availability of an adequate data set is a problem. For the second part of the process, a data set that contains data from public sources of a particular subject, together with whether they have committed fraud or not, is needed. At the time of writing, such a data set is unobtainable and impossible to put together. The first part requires a different data set, namely one that contains individual persons and their corresponding profiles on public sources. Although a data set that exactly matches these criteria has not yet been obtained, a test set could be constructed. This data set will be manually constructed in order to do some preliminary evaluation of the first part. However to fully evaluate both the first and second part, adequate data sets are necessary.

2.3 RESEARCH APPROACH

On a high-level overview, this research will exist of four phases that will be shortly described in this section. The first part of this research will be gathering data. A literature study focused on the following topics will be performed: business intelligence, web information extraction, entity matching and the legal issues. Additionally, exploratory interviews will be conducted. This first phase will result in artifacts from which the PSRA Process and PSRA Architecture will be constructed.

After — and based upon the results of — the first phase, an initial version of the PSRA Process and PSRA architecture will be constructed. Additionally, a comprehensive description of each of these artifacts will be written as well.

When the process and reference architecture are constructed, the next phase will focus on their evaluation. In order to evaluate them, a proof of concept will be implemented. Additionally, experts in the field will also evaluate both the PSRA Process and PSRA Architecture. From the evaluation of the proof of concept and the evaluation by the experts, a list of improvements for the architecture and the process will be created. The original process and reference architecture, adjusted with these improvements, will be the final versions of the PSRA Process and PSRA Architecture.

When both the process and reference architecture are finalized, the last phase consists of completing this thesis and delivering a presentation. These will result in the overall end-goal of this research project: graduation.

2.3.1 *Literature Review*

The literature review discussed in Chapter 3 is key to the overall research, as it provides the foundation for the process and architecture developed later on. Therefore, this section will purely focus on the ap-

proach taken to conduct the literature review. Chapter 3 will present the results of the literature review.

According to Webster and Watson (2002) a review of prior, relevant literature is an essential feature of any academic project. Since related literature will provide the foundation for the remainder of the research project, it is of great importance that the literature review should be conducted effectively. After all, the output of that literature review will be the input for the remainder of the study. In order to conduct the literature review effectively, it can be helpful to adopt a method. For this research, the literature review is conducted according to the method proposed by Levy and Ellis (2006).

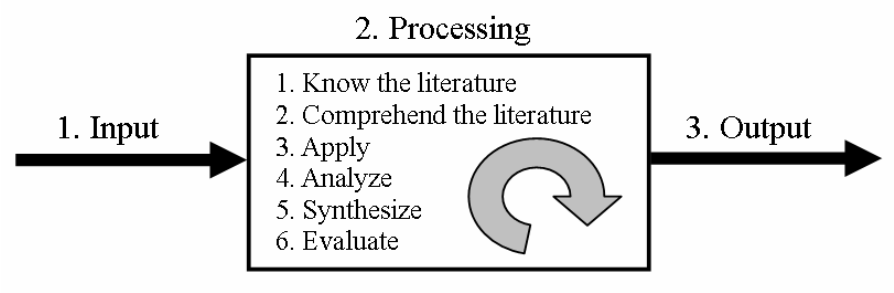

Figure 2.3.: The three stages of effective literature process (Levy & Ellis, 2006)

Levy and Ellis (2006) propose a systematic method for conducting and writing an effective literature review specifically for Information Systems (IS) research, consisting of three major stages: input, processing and output. As can be seen in Figure 2.3, the most elaborate stage is the central processing stage. However, Levy and Ellis (2006) note that the garbage-in/garbage-out problem applies to their method, since it is a systematic approach. This implies that, regardless of the quality of the central processing stage, the quality of the output will be low in the case of low quality input. Therefore, despite the fact that the central processing stage is the most elaborate, all stages are considered equally important during the literature review to ensure the overall quality.

Input

The first stage, the input stage, consists of literature gathering and screening. For this stage, they provide guidelines on three different aspects of gathering and screening literature. The first aspect involves how to gather quality literature, to prevent the garbage-in/garbage-out problem. In order to do so, Levy and Ellis (2006) propose to use only literature from high ranked journals, as these have been subjected to decent peer-reviews processes. Additionally, the use of conference proceedings should be limited as much as possible. This literature review follows these guidelines, as long as there is a sufficient amount of literature available in the high-ranked journals, if not then other, lower quality, sources are addressed.

The second aspect contains guidelines on how to actually find literature. Levy and Ellis (2006) recommend to use more than two different database vendors to find literature, thereby preventing narrowness of the literature review. However, specialised search engines

exist these days, such as Google Scholar. These make it possible to search multiple database vendors with only one search query, eliminating the need to access the multiple database vendors individually. Although this literature review slightly deviates from the guidelines by using these specialised search engines, basically it is still the same since Google Scholar certainly addresses more than two database vendors.

In addition, Levy and Ellis (2006) also provide three specific techniques - keyword searching, backward searching and forward searching - that can be used when conducting an effective literature review. Keyword searching is probably the most frequently used technique when conducting a literature review. In short, it refers to the use of a specific word or phrase in the query. However, according to Levy and Ellis (2006) this technique should only be used as a start, and not as the main step for a literature search. They note that identifying the first applicable keywords for a, to the researcher, unknown domain can be hard. This was indeed experienced during the literature review. It took some time before the Web Information Extraction and Entity Matching area were identified as keywords, partly due to the ambiguity in those domains.

The next technique, backward searching, builds on this initial keyword search. It consists of reviewing the references, and the earlier publications by the authors, of the literature found. Forward searching, on the other hand, consist of reviewing literature that has cited the literature found and the later publications by the authors of the literature found. All three techniques are used in this literature review, of which the latter two have become easier because of Google Scholar. Google Scholar provides easy access to the different techniques with functions as Related Articles, User Profiles and Cited by. Although this eased the process, manual backward and foreword searching is also done. Often this yielded different results than the Google Scholar functions.

The last aspect for which Levy and Ellis (2006) provide guidelines in the input stage, is how to identify when the literature search is finished. They describe this is as a feeling that arises, much like that your review is nearing completion when you are not finding new concepts in your article set (Webster & Watson, 2002). During the literature review this feeling only partly arose in the form that no new concepts were found *related to and in scope of the research*. In addition, on several occasions a lot of time went into understanding very detailed and specific literature, although this turned out to be too detailed or not directly relevant. In short, maintaining the scope of the research during the literature review was hard.

Processing

The processing stage consists of six sequential steps, in line with the six steps of cognitive development - knowledge, comprehension, application, analysis, synthesis and evaluation - of Blooms Taxonomy of Education Objectives (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Each successive step requires of more cognitively demanding activities. Although the processing stage seems to be the most elaborate, Levy and Ellis (2006) only provide techniques for two out of

the six steps. For the comprehension and application level they provide both guidelines and techniques, for the other four steps they only provide guidelines in the form of examples, explaining how to demonstrate that specific level mastery. These guidelines are adhered to as much as possible in the literature review.

Levy and Ellis (2006) provide two lists for the Comprehend the Literature stage, all related to key terminology often used by scholars. The first list consists of theories specifically used in the IS research field. Ditto for the second list that consists of constructs specifically used in the IS research field. These lists were not considered to be helpful during the literature review. Without explanation of each theory and/or construct it was hard to identify the ones that were applicable. Therefore, there was virtually no use of these lists during the processing stage of the literature review.

For the application level, a technique is provided based on the fact that a literature review is concept-centric (Webster & Watson, 2002). This implies that the literature will be grouped on concepts rather than chronological or on author. In order to effectively do this they advise to use a concept matrix, thereby identifying the concepts covered in the different articles. The concept matrix proved to be very useful during the literature review. It provided a clear overview of the literature identified thus far and allowed to easily identify whether literature contained new concepts or complemented existing concepts.

Output

For the final stage, the output stage, Levy and Ellis (2006) provide techniques and guidelines for writing arguments and for writing the literature review in general. For writing arguments they discuss two different, but relatively similar, argument theories. The major difference between the two models is the step wherein the claim is used. One argumentation theory uses the claim as first step, whereas the other uses it as the final step in the process. This literature review will use the latter, because it is consistent with the overall structure of the research. The literature is the first step, and provides the evidence and warrant for the final claims, which are the PSRA process and architecture.

To effectively write the literature review in general, they suggest a plan of action to write the review:

> "The plan should include pre-writing a literature review structure (i.e. an outline), allocating appropriate evidences for each section, developing the first draft, allocating appropriate time for revising the draft, and writing the final draft." (Levy & Ellis, 2006)

Although no explicit plan is created for this literature review, the steps mentioned have been taken. Especially pre-writing the outline of the literature review — the determination of the relevant research areas and their components — offered a good guidance during the further writing of the literature review.

Purpose

The purpose of this literature review consists of a threefold:

- Identifying artifacts (constructs, models, methods and instantiations) in the research areas relevant to this research, they will be combined to construct the PSRA process and architecture.
- Identifying related studies and indicate what is missing to solve the problem at hand.
- Fully understanding the context of the problem at hand and defining the related concepts within that context.

2.3.2 *Exploratory Interviews*

In order to get the current state of risk analyses in fraud sensitive business environments, three exploratory interviews with experts in the field were carried out. The purpose of the exploratory interviews are to get *a sense* of existing systems and the opinion of experts in the field on a system as proposed in this research, it is not a complete research into these matters. Although the three distinct experts all work for companies that operate in different sectors within the Netherlands, their daily tasks are all related to risk analyses, fraud detection and fraud prevention. Expert I is Manager Security at a top 3 e-tailer company, expert II is Fraud Coördinator at a top 3 credit provider and expert III is an analyst at the Inspectie Sociale Zaken en Werkgelegenheid (Ministry of Social Affairs and Employment). All of them have several years of experience related to risk analyses in fraud sensitive environments, often in multiple jobs and at multiple companies. An overview of the experts can be found in Table 2.1.

	Current		Previous	
#	Company	Function	Company	Function
I	Top 3 e-tailer	Manager Security (2 months)	Top 3 e-tailer	Manager Security (2 years)
II	Inspectie SZW	Analyst (7 years)	Centrale Justitiële Dienst	Analyst
III	Top 3 Credit provider	Fraud Coördinator (5 months)	Credit provider	Fraud Specialist (5 years)

Table 2.1.: Experts

An one-hour, semi structured interview was conducted with each of these experts in order to get their opinion and learn from their experience. After each interview a summary, which can be found in Appendix C, was written and sent to the expert for verification, along with the question whether this reflects their opinion and optionally some additional follow-up questions. As can be seen in the interview protocol in Appendix B, the following topics were addressed: fraud as it relates to the company, manual measures to detect and prevent fraud, automated systems to detect and prevent fraud, the manual use of personal data from public sources, the automated use of personal data from public source, and ethical issues. The experiences

and opinions of the experts on these topics are taken together and addressed in Section 4.

2.3.3 *Expert feedback*

In addition to the evaluation through a proof of concept, two experts were also asked to give feedback on the developed process and architecture. This has been done in order to slightly compensate for the fact that not all parts of the process and architecture were evaluated through the proof of concept, since no cooperative organization was found for this research project. Two experts, active within the Business Intelligence field, have been asked to evaluate the artifacts. More information on the experts can be found in Table 2.2.

Name	Company	Function
Hans Geurtsen	Info Support	Business Intelligence Architect
Koos van Strien	Info Support	Business Intelligence Consultant

Table 2.2.: Expert feedback experts

The PSRA Process and PSRA Architecture have been sent to the experts, with the accompanying textual explanation. The experts were asked the following to look closely at each step of the process and each part of the architecture and determine from their experience if this step or part makes sense. Additionally, they were also asked if they missed any crucial steps or parts in the process and architecture. Some of their feedback is used to improve the results presented in this research itself, the rest is discussed — together with the improvements identified by the researcher — in Chapter 7.

Part II

RESULTS

3 LITERATURE REVIEW

This chapter will present the results of the literature review conductd. First, the concepts central in this research - business intelligence, web information extraction and entity matching will be defined and introduced in separate sections. Hereafter, a section will be dedicated to the legal issues.

3.1 BUSINESS INTELLIGENCE

3.1.1 *Definition*

The term Business Intelligence was coined by Luhn in 1958. In his article "A Business Intelligence System" he describes a system that has many remarkable similarities to what we consider Business Intelligence systems today. In his article he defined the term Business Intelligence as follows:

> "Business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal." (Luhn, 1958)

In short, according to Luhn (1958) a business intelligence system should facilitate the communication within a business, ensuring that the right information is delivered to the right location within that business.

A few decades later, in 1989, Business Intelligence was made more wide-spread by Howard Dresner, who defined it as "a set of concepts and methods to improve business decision making by using fact-based support systems" (Power, 2007). Although the importance of facts in Business Intelligence was already indicated by Luhn (1958), Howard Dresner added the notion that Business Intelligence should support the decision making process within a business. This addition has proven to be an important part of Business Intelligence to this day, and most modern definitions still, either explicitly or implicitly, include this aspect.

Since then the term and the research area, Business Intelligence, have become more mature. The definition of Turban et al. (2012) is used for this research, they define Business Intelligence as "a conceptual framework for decision support. It is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies". This definition positions Business Intelligence more as a separate area, containing various elements. This research

will utilise some of these elements, and combine them with elements from other areas.

From Negash (2004) it gets apparent how close Luhn (1958) was, although not explicitly from the definition, but from the following paragraph: "Implicit [..] is the idea (perhaps the ideal) that business intelligence systems provide *actionable information delivered* at the right time, *at the right location*, and in the right form to assist decision makers" (Negash, 2004). Something Luhn (1958) noticed almost 50 years earlier.

3.1.2 *Architecture*

Contrary to the controversy regarding the Business Intelligence definition is the consensus regarding the high level architecture of Business Intelligence systems. Figure 3.1 to Figure 3.6 depict several architectures published between 1997 and 2012 in articles and books. Although some differences can be identified on the surface, this is mainly due to the choices of the author(s) related to naming and visualising their concept. Apart from a few less-essential differences, they consist of the same components.

The first component, *data sources*, is depicted in all five architectures. Some authors explicitly distinguish between external and internal sources (Chaudhuri & Dayal, 1997; Chaudhuri et al., 2011; Turban et al., 2012) with which they implicitly indicate that external sources are considered an important part of Business Intelligence systems. Although only a few authors mention web pages as an external data source (Chaudhuri & Dayal, 1997; Chaudhuri et al., 2011) in their accompanying text, it is evident that these fit in the category of external sources. This research will also take this division into account by introducing a separate Extract, Transform and Load process for web pages.

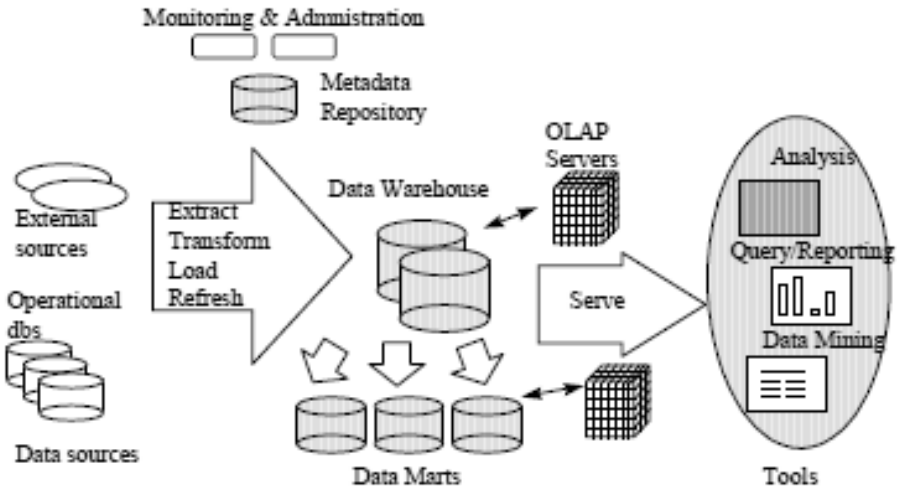

Figure 3.1.: Business Intelligence architecture adopted from Chaudhuri and Dayal (1997)

The second component, the *Extract, Transform and Load* process is also depicted in most architectures. Only Negash (2004) does not explicitly depict the Extract, Transform and Load process but he does describe it in his accompanying text. As previously mentioned, this

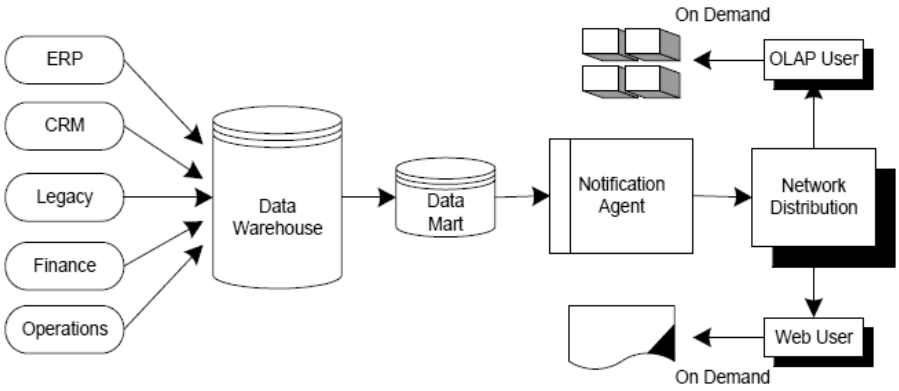

Figure 3.2.: Business Intelligence architecture adopted from Negash (2004)

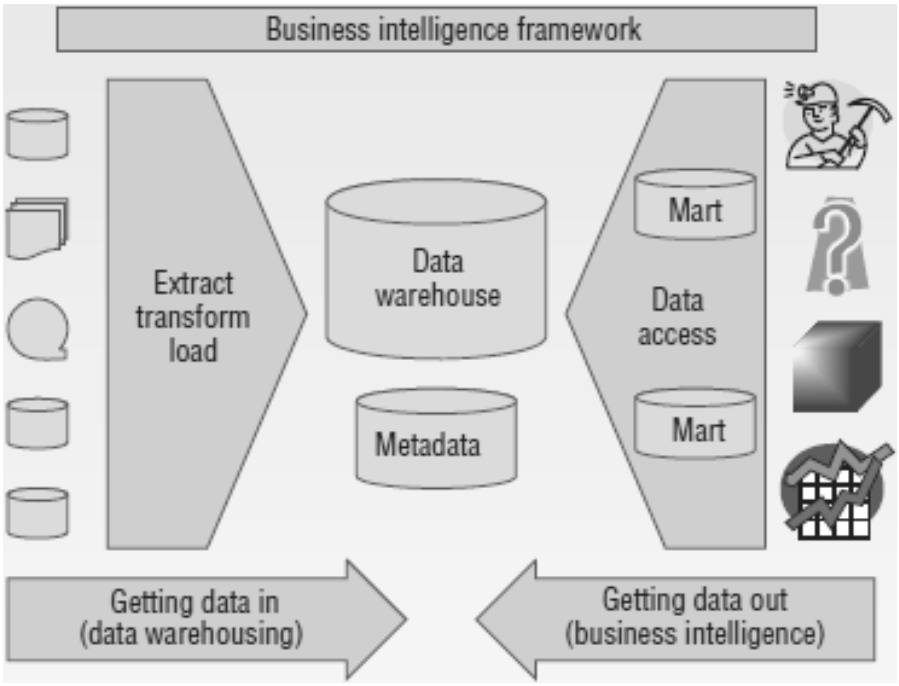

Figure 3.3.: Business Intelligence architecture adopted from Watson and Wixom (2007)

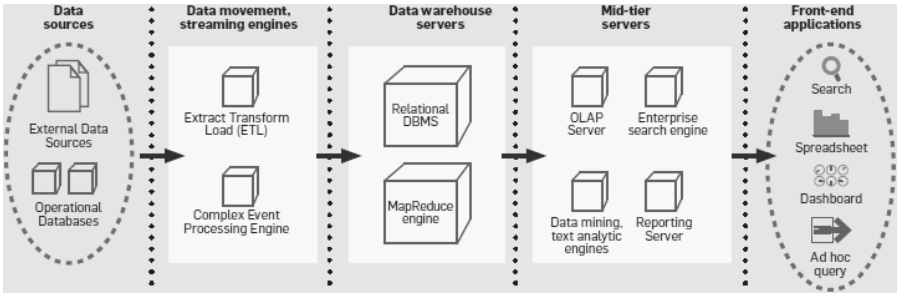

Figure 3.4.: Business Intelligence architecture adopted from Chaudhuri et al. (2011)

research will utilize a specific approach for extracting personal data from web sources, which will be addressed in Section 3.1.3 and Section 3.2.

The third and last component within the scope of this research, is the *data warehouse*. Two approaches are visible in the depicted architectures, namely the approach with data marts (Negash, 2004; Watson

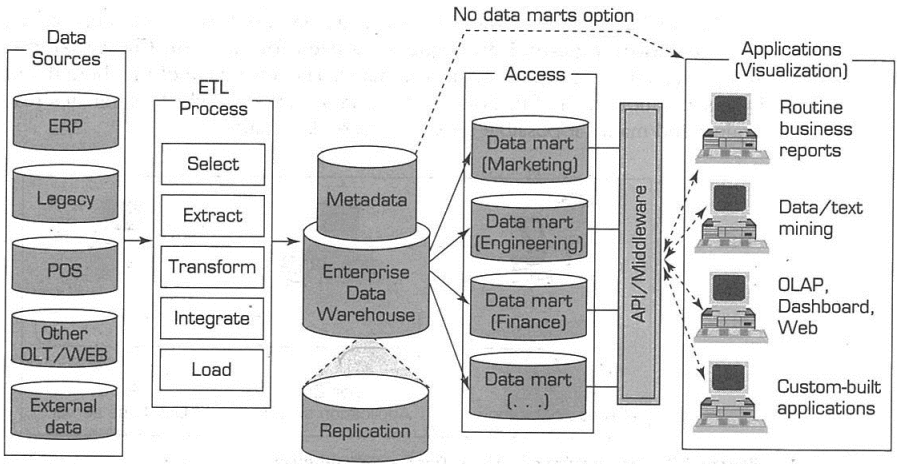

Figure 3.5.: Business Intelligence architecture adopted from Turban et al. (2012)

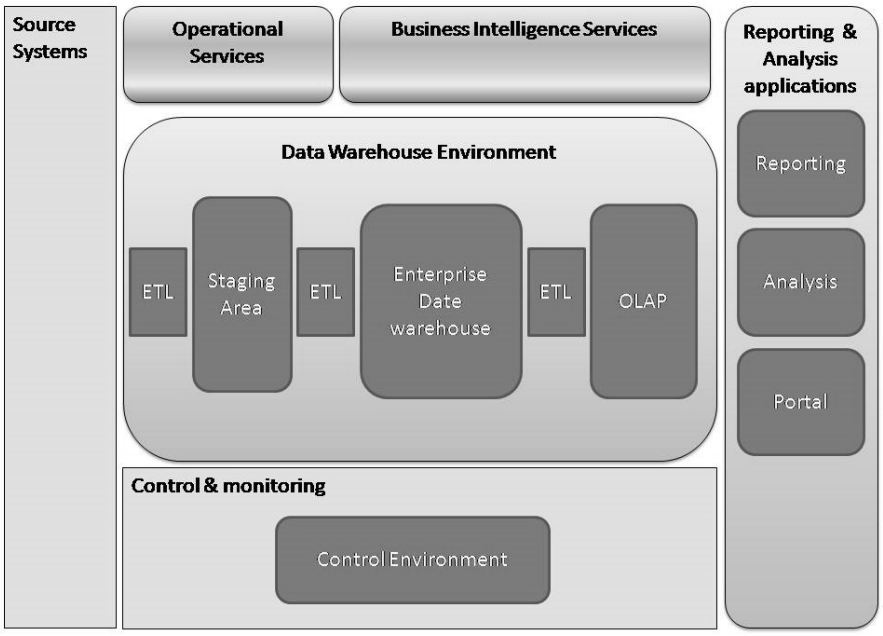

Figure 3.6.: Business Intelligence architecture adopted from Info Support

& Wixom, 2007) and the approach without data marts (Chaudhuri et al., 2011). Some authors integrate both possibilities in their architecture by including a "no data marts option" (Turban et al., 2012) or by adding a direct relation between the data warehouse and tools as well as between the data marts and the tools (Chaudhuri & Dayal, 1997).

The remaining components, which are outside the scope of this research, can be grouped under tools and applications, and servers designated to support these tools and applications. Because this study focuses on extending current risk assessments systems by including personal data from web sources, these tools, applications and servers are considered to be already available or left for future research.

In the preceding section it was noted that the Business Intelligence System described by Luhn in 1958 showed many remarkable similarities with contemporary Business Intelligence systems. His Business Intelligence system had both internal and external documents

(data sources) from which information was extracted with the use of auto-encoding and auto-abstracting (extract, transform and load). This was then stored in a microcopy storage (data warehouse) and matched to the people for whom the information was interesting (analysis tool). It even included a desk print and a display screen to present the information to the end-user (application). Luhn was years ahead of his time.

From the comparison of the architectures in this section the first artifact is extracted in the form of a high level architecture of a system that the PSRA process implements. The artifact is depicted in Figure 3.7. It consist of the web sources from which the data is extracted, transformed and loaded into the data warehouse. It should be noted that this figure only depicts the parts of the architectures within the scope of this research as discussed in Section 2.2. Therefore, the remaining components — such as tools and applications — are not depicted in this figure since they do not apply to this research. The next and subsequent sections will zoom in on the extract, transform, load process and the data warehouse, respectively.

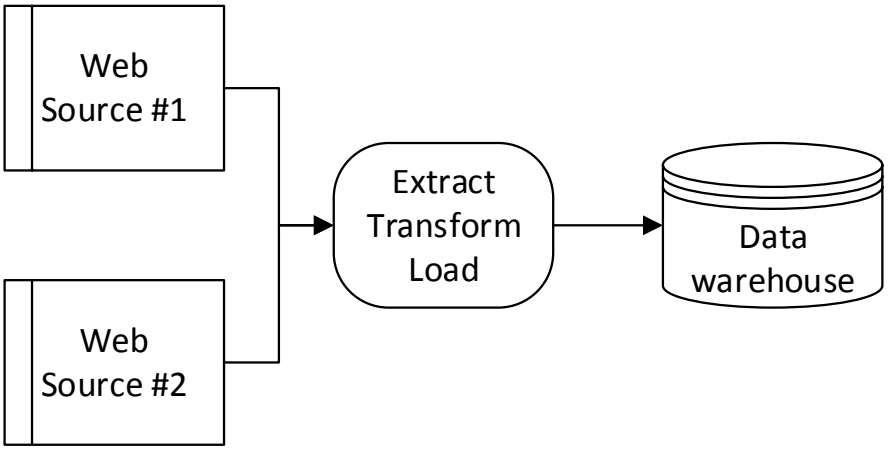

Figure 3.7.: High-level architecture of the "getting data in" part of a Business Intelligence system with web sources

3.1.3 *Extraction, Transformation and Load*

The extract, transform and load process is "a data warehousing process that consists of extraction (i.e., reading data from a database), transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and load (i.e., putting the data into the data warehouse)" (Turban et al., 2012). A point of criticism on this definition is that data is not always extracted from a database, it can also be extracted from other kinds of sources. As in this study, for example, reading data from a web source. However, when extracting personal data from web sources three additional problems arise that should be addressed.

The first problem arises from the difference that the Internet primarily consists of unstructured and semi-structured data (Blumberg & Atre, 2003; Embley et al., 1999) and analysis tools, such as data mining tools, operate with structured data from the data warehouse.

In order to solve this problem, structured data should be extracted from unstructured and semi-structured data so that it can be loaded into the data warehouse.

The second and third problem are related to the fact that a system implemented with the PSRA process will utilise personal data from web sources. On one hand, web sources that contain personal data, such as Facebook, contain too much data about to many different people in order to extract every single piece of data from all Facebook profiles. Especially for this research it is more efficient to only extract data from Facebook profiles that at least show some similarity in the person name with the subject of the risk assessment. On the other hand, person names on the Internet are highly ambiguous (Artiles, Sekine, & Gonzalo, 2008). Different people share the same name, for example *James Smith* or *Bas Jansen*, making it hard to retrieve the personal data that actually relates to the subject of the risk assessment.

The problem of extracting structured data from unstructured or semi-structured data sources is central in the Web Information Extraction research area, and will be addressed in Section 3.2. Solving the problem related to the highly ambiguous person names on the Internet is the focus of the entity matching research area, addressed in Section 3.3. These two research areas will substantiate the extraction of data in the extract, transform and load process as seen in Figure 3.8. The transformation and loading of the data into the data warehouse does not differ from the normal extract, transform and load process once the structured data is extracted from the unstructured and semi structured data.

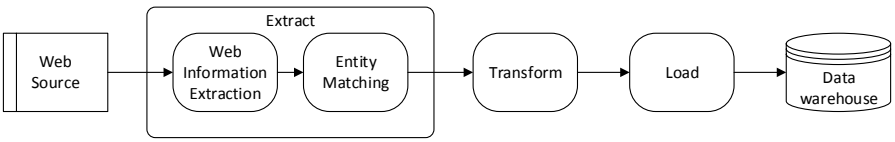

Figure 3.8.: Extract, transform, load process for personal data from web sources

3.1.4 *Data warehouse*

According to Inmon (2005) a data warehouse is "a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process". Subject-oriented constitutes that the data within the data warehouse is organized by subjects, such as sales, products or customers. Integrated refers to the diverse data that is extracted from different sources and integrated into one consistent format in the data warehouse. The data warehouse contains historical data, which creates a time-variant collection of data. Once data is entered into the data warehouse it normally is not changed or deleted, making the collection of data non-volatile. The definition is in line with the earlier discussed definition of business intelligence, as it supports the decision making process of the management.

Besides Inmon (2005), another giant in the field is Kimball (Breslin, 2004). Kimball (2006) defined a data warehouse as "a copy of transaction data specifically structured for query and analysis". It is evident that this definition is less detailed then the one by Inmon (2005) and thus less specific. He omits that a data warehouse is used to support decision making, and he also omits how the data should be structured within the data warehouse. Besides the differences that become clear from the definitions, they also both have significantly different ideas about the architecture and the methodology of a data warehouse (Breslin, 2004). However, this discussion is out of scope of this research, interested readers are recommended to read Breslin (2004) for a full discussion. This research assumes that an existing Business Intelligence system is already in place, and hence the data warehouse as well. In addition, the proposed solution in this research does not depend on the type of data warehouse since it focuses on the extract transform and load process.

3.2 WEB INFORMATION EXTRACTION

Although Web Information Extraction finds it origin in Information Extraction, the task at hand differs largely from the traditional Information Extraction task (Chang, Kayed, Girgis, & Shaalan, 2006). An Information Extraction task consists of extracting information from a given input into a extraction target. Originally, the input for Information Extraction tasks were primarily unstructured documents. The extraction target "can be a relation of k-tuple (where k is the number of attributes in a record) or it can be a complex object with hierarchically organised data" (Chang et al., 2006). For instance, a social network profile may include only attributes such as name, age and gender. However, when it also allows persons to indicate a basically unlimited list of favourite movies, it becomes a complex object instead of a simple relation of k-tuple.

The large difference between the Web Information Extraction and the traditional Information Extraction task is due to the difference in the type of input (Chang et al., 2006). On the one hand, the traditional Information Extraction task is focused on unstructured free texts. On the other hand, the Web Information Extraction task focuses on semi-structured web pages. Because of this difference in the input type, other techniques are also required in order to successfully extract information from the input. Normally, for the traditional task techniques from the Neuro-linguistic programming research area —- an area focussed on are adopted, whereas the web task mainly relies on extraction rules (Gregg & Walczak, 2006).

For Web Information Extraction, so-called wrappers are put in action to do the actual extraction (Chang et al., 2006; Gregg & Walczak, 2006; Laender, Ribeiro-Neto, da Silva, & Teixeira, 2002). These wrappers are, generally, site-specific programs that understand the information that resides within the semi-structured web pages (input) and is able to extract this information as structured data (output). Wrappers will be the subject of the following section.

Web Information Extraction will be the first link in the chain of the extraction phase from the extract, transform, load process as dis-

cussed in Section 3.1.3. Figure 3.9 visualises the knowledge from the Web Information Extraction research area that will substantiate a part of the PSRA method. It depicts an architecture to extract information from web sources. Multiple web sources are visible on the left-hand side along with a separate wrapper specifically developed for each one of them, although only two are displayed this can be expanded to a theoretically infinite amount. These wrappers are created with, or with the aid of, wrapper creation tools. Depending on the wrapper creation tool a certain degree of automation is achieved, ranging from a little guidance by specific wrapper programming languages to full automation by ontology-based extraction tools. At the right-hand side of the diagram the arrows coming from the wrapper resemble the extracted data that is passed on to the next link in the extraction chain, the entity matching, which will be discussed in Section 3.3.

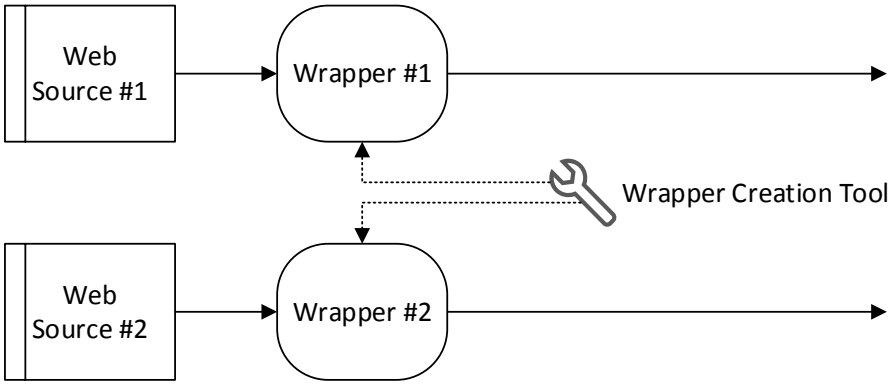

Figure 3.9.: Web Information Extraction architecture

3.2.1 *Wrappers*

Traditionally, wrappers were hand-coded for each specific site with general purpose languages, making the creation of wrappers a costly task. Since then, a lot of research has been conducted to make this easier. The first efforts were the development of programming languages, specifically designed for extracting information from web pages. These made the task of hand-coding wrappers somewhat easier, but there was still a great deal of work involved. Hereafter, multiple techniques were developed that made the creation of wrappers much easier. Gregg and Walczak (2006) state that two different approaches can be identified at a high-level, namely position-based extraction and ontology-based extraction.

Ontology-based extraction offers the highest automation degree of the two, it requires an one time investment to built the ontology where after an extraction tool can utilise this ontology to extract information from various websites, even if the web source changes the HTML-structure over time. However, building an ontology is an entire area of research in itself and it is difficult to capture a complete domain. Additionally, according to Gregg and Walczak (2006) not all web data is suitable for ontology-based extraction as it lacks unique characteristics or keyword labels. Thus although ontology-based offers the highest automation degree and resilience, it is relatively com-

plex to implement, especially when other more simple tools are perfectly suited for the task at hand.

Tools that utilise position-based extraction rely on the structural features of the web page. It uses extraction rules that indicate on what position certain data can be found within the document. When web pages are generated and filled with data from a database, it means that detail pages for multiple items share the same structural features, and thus the same extraction rules can be used. Therefore, it is only necessary to determine the extraction rules for an item on a specific web source once, which can then be reused to extract information from all items on that specific source. This is also the case with, for example, Facebook. All profiles are generated using the same template, only the interesting data differs among the profile and should be able to be extracted in the same manner on all profiles.

3.2.2 *Automation*

The latter category, tools based on position-based extraction, utilize machine learning techniques to automate the creation of the extraction rules used in the wrappers (Chang et al., 2006). According to Chang et al. (2006) three different degrees of automation in generating these wrappers exist, namely supervised, semi-supervised and unsupervised. Tools adopting the supervised approach require a complete and exact set of example web pages from the source labelled by a user. Based on this set the machine learning algorithm determines the extraction rules and therewith creates the wrapper for the specific source. The semi-supervised approach only requires a rough set of example web pages labelled by a user. Finally, the unsupervised approach does not require any of the example web pages to be labelled. Instead it automatically extracts potentially interesting information from the web pages. However, some post-processing may be required, for example to select only he relevant data or assign labels to extracted data (Chang et al., 2006).

3.2.3 *Levels*

Sarawagi (2002) distinguishes three types of wrappers, namely record-level, page-level and site-level wrappers. Record-level wrappers extract a list of homogeneous records that reside on one page, such as the individual profiles on the search results pages on Facebook which is displayed when one searches for on specific personal name. Page-level wrappers extract different kinds of records residing on one page, for instance check-ins, posts and personal information on a Facebook profile page. Finally, site-level wrappers extract information from multiple pages of a web site, thereby constructing a database with all the desired data from the web site. For the purpose of the PSRA method only record-level and page-level wrappers are needed, a requirement which is also apparent from the application context.

As discussed in Section 3.1.3 a web source such as Facebook contains a lot of data about many different people. In order to extract every single piece of data from all Facebook profiles a site-level wrapper would need to be created. But, only a very small percentage of

the profiles is actually useful for risk analyses, as the subjects of the risk analyses are only a small percentage of the world population. Therefore, it is more efficient to use the search engine provided by those website to identify profiles that at least show some similarity with the subject of the risk assessment. This can be accomplished by creating a record-level wrapper that wraps the search engine result from the web source. In addition, data should also be extracted from the profile page itself, therefore a page-level wrapper is needed that wraps the profile page.

These findings are visually depicted in Figure 3.10 that displays the architecture of a wrapper that is able to extract information from web sources containing personal data. The search results wrapper is a record-level wrapper that is able to extract information from the search results page of the web source given a specific query. The first x items are extracted and passed on to the item wrapper, which is a page level wrapper that is able to extract all the data residing in the items detail page. This data is then passed on to the next link in the chain, the Entity Matching, which will be discussed in the next section.

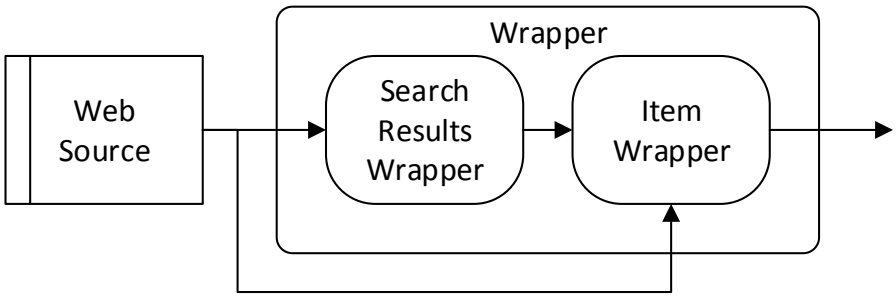

Figure 3.10.: Wrapper architecture with separate record-level and page-level modules

3.3 ENTITY MATCHING

3.3.1 *Task definition*

Entity matching (also referred to as entity linkage, entity resolution, entity identification, entity consolidation and entity disambiguation) is "the task of identifying entities referring to the same real-world entity" (Köpcke & Rahm, 2010). By using the general term entity, the task can be considered very broad. More specialised research areas exist for specific entities, such as person matching for entities of the type person. Nevertheless, mostly the same techniques are used among the different specialised research areas. Although the task at hand slightly differs (this will be addressed shortly) from the task described in this definition of entity matching, the same techniques can be applied to solve it.

The entity matching problem was first formalized by Fellegi and Sunter in 1969, they developed a mathematical model that encompassed the entity matching problem. The problem starts with two sets of entities A and B of which all pairs of entities that represent the

same real-word entity should be identified. In order to accomplish this, first all possible entity pairs from set A and set B are considered.

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

For each of these entity pairs it should be determined whether they represent the same real-world enity or not. This is formalized by defining two new sets, M for matched pairs and U for unmatching pairs.

$$M = \{(a, b) : a = b, a \in A, b \in B\}$$

$$U = \{(a, b) : a \neq b, a \in A, b \in B\}$$

The task referred to in the definition by Köpcke and Rahm (2010) is the task of deciding for each pair to which of the two sets they belong.

As previously discussed the task at hand in this research slightly deviates from the defined entity matching task. The problem in this research starts with a set of persons (entities) X, originating from an internal database. For a person $x \in X$ a subset Y of profiles that potentially represent the same person is extracted from the entire set of profiles Z within a specific web source.

$$Y \subseteq Z = \{z : z \sim x, Z \in Z\}$$

The task is then to decide which profile z in subset Y represents the same real-world person as x. This task should be repeated for each web source that should be included, thereby selecting one (or none) profile per web source per person x. Of course, this task should be repeated for each person in X.

3.3.2 *Solution*

Although the type of task slightly differs, at a lower level the problems that should be solved are essentially the same. In both situations it has to be decided whether the entities in a specific pair represent the same real-world entity or not, regardless of the fact how the pair is established. This is done by a decision function (also referred to as decision model and matching model) "that makes the decision of whether a record pair is a match, non-match or possible match." (Gu, Baxter, Vickers, & Rainsford, 2003). In order to be able to make this decision, the decision function utilizes one or more similarity functions (also referred to as similarity metrics, matchers and comparator functions).

According to Köpcke and Rahm (2010) three approaches for the decision functions can be distinguished: numerical, rule-based and workflow-based approaches. The numerical approach combines outcome of the individual similarity functions numerically, for example by taking average or weighted average of the individual similarity values computed on the specific attributes. The rule-based approach uses rules to determine whether two entities are similar or not. An example rule could be that two entities are considered the same person when the name and city attribute between both entities exceed a certain threshold. Finally, the work-flow based approach makes it possible to define a particular sequence wherein similarity functions

should be executed, which improves the matching with each subsequent step.

For the similarity functions two approaches are considered, attribute value similarity functions and context similarity functions (Köpcke & Rahm, 2010). Although Köpcke and Rahm (2010) mention context similarity functions as a type, other surveys, such as Gu et al. (2003) and Elmagarmid, Ipeirotis, and Verykios (2007), only mention attribute value similarity functions.

The attribute value approach uses, as the name implies, the value of an attribute describing the entities. The function compares the same attributes of both entities in the entity pair and computes a similarity value. This similarity value resembles the similarity between the two entities on a specific attribute and usually ranges from 0 to 1 (Köpcke & Rahm, 2010; Gu et al., 2003). Most attribute value similarity functions are actually string similarity functions and there exist different variations of those.

An example of a string similarity function is the Levensthein distance proposed by Levenshtein in 1966. The Levensthein distance between two strings is determined by the number of operations that is needed to transform one string into the other string. The allowed operations are insertion, deletion and reversal and the minimum of amount of operations with which the transformation is achieved is the Levensthein distance. So for the surnames "Janssen" and "Jansen" the Levensthein distance would be 1 since only a s has to be inserted (or deleted). Since this represents a distance value, and rather a "closeness" value is needed, the value needs to be translated. Additionally, it also needs to be normalized in order to have it range from 0 to 1. To normalize the value, it is possible to divide it by the maximum Levensthein distance (the length of the longest word). This resulting value is subtracted from 1 in order to translate it to a closeness value. This is formalized in Algorithm 1.

Algorithm 1 Normalizing the levensthein distance

```
1: function NORMALIZEDLEVENSTHEIN(String string1, String string2)
2:     levensthein ← levensthein(string1, string2)
3:     maxLength ← string1.length
4:     if maxLength < string2.length then
5:         maxLength ← string2.length
6:     if maxLength = 0 then
7:         return 1
8:     else
9:         return 1 - (levensthein / maxLength)
```

In the example of "Janssen" and "Jansen" the Levenshtein distance is 1 and the length of the longest word is 7, so the normalized Levenshtein similarity value is:

$$1 - (1/7) \approx 0.86$$

Of course, other string similarity functions exist, and those might be more useful depending on the dataset.

On the other hand, the context similarity functions use information about the context that can be mapped to a graph to compute the

similarity. So instead of using information that describes the entity itself, it utilizes information that describes the relation with other entities. When matching profiles from web sources this approach could potentially be very useful as they often include connections between friends, for example on Facebook, that can be easily mapped to a graph. However, in the use case supported by this research it is required that this graph can also be constructed from the internal database in order to compute the similarity. Unfortunately, this type of information is usually not present in internal databases. As in the Facebook example, very few companies will have the network of friends of a person internally present.

Which of the three decision function approaches, similarity functions and attributes should be used with a particular data set is unfortunately a question that is unlikely to be resolved soon: "The duplicate record detection task is highly data-dependent and it is unclear if we will ever see a technique dominating all others across all data sets" (Elmagarmid et al., 2007). Therefore, choosing the right similarity functions, attributes and approach for the decision function can both differ among the selected public sources and depend on the the internal data set as well. Thus determining these in advance is impossible and it will therefore be a step in the PSRA process. As an example, above it was discussed that context similarity values are less suitable because often the internal data set is lacking the needed information. However, if a company implements the system proposed in this research does have the needed information internally available it does become an option. This is something that should be examined and decided by each implementation.

3.3.3 *Entity Matching Frameworks*

The preceding problem has not yet been solved in current research, but another problem related to choosing the right entity matching strategy has been solved. Apart from choosing the similarity functions, their attributes and an approach for the decision function, it should also be decided how the decision function exactly combines the output of the different similarity functions in order to make a decision. For example, when the decision function uses a numerical approach in which each output gets assigned a weight, an attribute which is more decisive for the overall similarity, such as the name, should get assigned a greater weight than an attribute that is less decisive such as the province. In determining the best distribution of these weights, training-based entity matching frameworks can assist as they "optimize the combination of a manually predetermined set of matchers" (Köpcke & Rahm, 2010). For clarity, they are not yet able to determine the attributes, similarity functions and the approach for the decision function and thus these are considered predetermined.

These training-based frameworks assist in three different degrees of automation: manual, semi-automatic and automatic (Köpcke & Rahm, 2010). This closely resembles the three degrees of automation discussed in 3.2.2 unsupervised, semi-supervised and unsupervised, respectively. Manual requires the user to choose entity pairs and label them by hand according to whether they match or do not

match. Based upon these labeled entity pairs the framework determines the optimal matching strategy. This is already an improvement compared to manually determining the optimal matching strategy because choosing and labeling entity pairs is an easier task. Semi-automatic frameworks only require some entity pairs to be labeled by the user, and these are often proposed by the framework (active learning). Lastly, automatic training-based entity matching frameworks perform both the task of choosing the entity pairs and labeling them. Although these training-based frameworks ease the task of determining the optimal matching strategy, they do require an adequate training data set (Gu et al., 2003; Elmagarmid et al., 2007; Köpcke & Rahm, 2010). If such an adequate training data set is not present, the matching strategy will have to be determined manually.

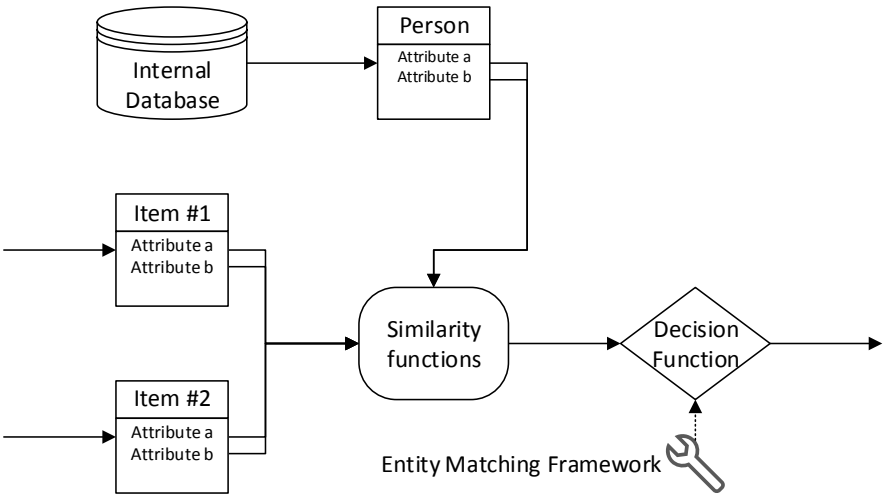

Figure 3.11.: Entity matching for profiles on public sources

Figure 3.11 visually depicts the entity matching concepts as a model. At the core of this model are the similarity functions that compute the similarity between each person in the internal database and each item originating from the web information extraction link. This is done for each attribute that is included in both entities, or based on contextual information when operating context similarity functions. The decision function utilizes the output of the individual similarity functions and combines them in order to make a decision, therewith picking one (or none) profile that refers to the same real-world entity as the person record. This entire process is repeated for each web source since the similarity functions and decision function can differ due to the fact that the data set from the web source is different (it might contain more, less or other attributes). The entity matching framework optionally assists the user in determining the parameters of decision function, but not the type of decision functions. Probably, the entity matching framework will play a greater role in the future by assisting in determining the type of decision function and the similarity functions.

3.4 RELATED STUDIES

To answer some of the research questions of this research the PRSA process and architecture are proposed, which are built by combining

	Approach	Intended context	Automation	Implementation
Pawar & Sharda, 1997	Framework	Strategic decision making	Manual	Standalone
Srivastava & Cooley, 2003	Architecture	Organisation decision making for competitive advantage	Semi-automatic	Standalone
Soper, 2005	Architecture	None	Automatic*	Extension
Baumgartner et al., 2005	Tool	None	Semi-automatic	Extension
This	Process and architecture	Decision making based on personal information	Semi-automatic	Extension

Table 3.1.: Comparison of related studies and this research

artifacts identified in different applicable research areas. Each applicable research area has been, in turn, a source for a lot of related literature. However, some studies already partly took this approach as well. This section will address these studies and discuss not only the similarities, but also the differences that distinguish this research from the existing studies. Most differences and similarities are summarised in Table 3.1.

Several studies exist that have investigated the combination of web sources and Business Intelligence. One of the earliest mentions of this combination is by Pawar and Sharda (1997), they propose a framework that assists in using web sources to obtain information used for strategic decision making. Srivastava and Cooley (2003) even coined the term Web Business Intelligence for the "emerging class of software that leverages the unprecedented content on the Web to extract actionable knowledge in an organisation settings". In addition, they present a high-level architecture for Web Business Intelligence systems and discuss which technologies can be utilised for the components within that architecture. Soper (2005) build upon this research and propose an architecture that guides the development of so-called Automated Web Business Intelligence systems, which he defines as "software applications that utilise automated processes in order to extract actionable organisational knowledge by leveraging the content of the web". Finally, Baumgartner et al. (2005) present a tool called Lixto that is able to automatically extract information from web sources and then transform this information for use in Business Intelligence systems. Many tools that automatically extract information from web sources exists these days, but Baumgartner et al. (2005) were one of the few that described how it could be combined with Business Intelligence systems.

Although these studies, including this research, all share the same common goal to support decision making by utilising information from web sources, there are also major differences among themselves and in comparison with this study. First, the approach to reach that

goal differs greatly. Pawar and Sharda (1997) propose a framework that guides practitioners in using web sources for the acquisition of external information. Both Srivastava and Cooley (2003) and Soper (2005) propose an architecture for systems that are able to support decision making with information from web sources, meant for guiding the implementation of these systems. Baumgartner et al. (2005) reached the goal by developing a commercial tool that actually performs the task of extracting information from web sources for Business Intelligence systems. This research will mainly focus on developing an architecture and a process to guide the implementation of such systems. In addition a prototype will be built to validate the architecture and process.

A second difference is the context wherein the proposed solutions are intended to be used. In contrast to Srivastava and Cooley (2003) and Baumgartner et al. (2005) who propose their solution for use in all areas of decision making, Pawar and Sharda (1997) and Soper (2005) propose a solution that only supports strategic decision making and organizational decision making for competitive advantage, respectively. The solution proposed in this research is in line with the latter category, focusing specific on decision making that benefits from utilising personal information from web sources. Because of this specific intended context new problems arise (as discussed in Section 3.1.3) that should be solved and therefore the content of this research is fundamentally different.

Third, the studies differentiate in the degree of automation that the proposed solutions offer. Only Soper (2005) indicate that their solution is completely automated and do not allow for manual data gathering from the web. Indeed, no data is manually gathered from the web within their solution, however their solution is not totally automated. During initiation the systems are "initially provided with a single web data source from which to gather information" instead of finding this source on their own. Additionally, involvement during run time is also necessary because in order "to accurately determine the contextual relevance of a candidate web data source [..] a manual confirmation mechanism may be necessary". The framework proposed by Pawar and Sharda (1997) offers no automation at all, this is because at the time of publication Business Intelligence was a role within the company carried out by one or more employees instead of an information system as we know nowadays, therefore it only provides guidance to the employees during their work instead of replacing them. In between are the solutions proposed by Srivastava and Cooley (2003) and Baumgartner et al. (2005), which can be considered semi-automated. Although the actual extraction of the information from web sources is automated, the wrappers used for the extraction should be created beforehand. Srivastava and Cooley (2003) indicate that manually developing these wrappers is very intensive and recommend to use tools that generate these wrappers. The Lixto tool can be used to overcome the problem of manually developing wrappers, it provides the user with a graphical user interface to easily create wrappers. The PSRA process will also recommend to use tools that generate wrappers whenever possible just like Srivastava and Cooley (2003), thereby it can be considered semi-automatic.

The fourth, and final, point in which the solutions differ is whether the proposed solutions extend current systems or are intended to be standalone systems. Soper (2005) clearly state that their solution is "intended to provide a supplementary source of decision support information that can be integrated into an existing organisational decision making infrastructure", and Baumgartner et al. (2005) present their tool as an extension to current Business Intelligence systems that only utilise internal sources. Pawar and Sharda (1997) and Srivastava and Cooley (2003) propose their solutions as separate systems that should be implemented as a whole, this is also evident from the fact that they do not mention internal sources in their solutions. The solution proposed in this research also extends current systems, namely customer risk analyses systems.

In summary, this research shares the same goal as the other related studies, namely supporting the decision making by utilising information from web sources. The specific context, however, differs from the existing studies. Because of this specific context new problems arise (see Section 3.1.3), problems that have not yet been addressed in this context in scientific literature. The degree of automation and the type of implementation are semi-automatic and extending an existing system, respectively. In addition, none of the related studies addresses legal issues and only one study addresses ethical issues by discussing privacy matters (Srivastava & Cooley, 2003). This research will try to fill these gaps by providing solutions for the problems within the specific context and by addressing both legal and ethical issues.

3.5 LEGAL ISSUES

3.5.1 *Constitution*

In the Netherlands, privacy of the citizens is considered a fundamental principle. This is evident from the fact that the right to respect for their privacy is part of the "*Grondwet voor het Koninkrijk der Nederlanden* (2008)", which is the constitution of The Netherlands. Article 10 from the English version (*The Constitution of the Kingdom of the Netherlands*, 2008) is as follows:

> 1. Everyone shall have the right to respect for his privacy, without prejudice to restrictions laid down by or pursuant to Act of Parliament.
> 2. Rules to protect privacy shall be laid down by Act of Parliament in connection with the recording and dissemination of personal data.
> 3. Rules concerning the rights of persons to be informed of data recorded concerning them and of the use that is made thereof, and to have such data corrected shall be laid down by Act of Parliament.

Because the right to respect for their privacy is considered a fundamental principle, it is almost inevitable that legal implications arise when personal data is processed. The second and third paragraph of the same article indicate more implications ahead, since these paragraphs regulate that there shall be rules concerned with personal

data, laid down by Act of Parliament. The rules mentioned in these two paragraphs are defined in a separate Act, namely the "*Wet bescherming persoonsgegevens* (2001)", which will be addressed shortly.

In order to put this in a — more general — international context, the European Commission for Democracy through Law maintains a database which systematically indexes the constitutions of numerous countries. This database is publicly available on `www.codices.coe.int` and allows anyone to effectively search for parts of the constitution. By using the indexes displayed in Table 3.2, implementers of the PRSA system can search for relevant parts of the constitution applicable in their country. Therewith it is possible to get informed about the legal implications that could arise in a specific county during the execution of the PRSA process.

Index	Description
5.3.32	Right to private life
5.3.32.1	Protection of personal data
5.3.33	Right to family life

Table 3.2.: Relevant indexes on `www.codices.coe.int`

3.5.2 *Personal Data Protection Act*

The above-mentioned, more specific, act '*Wet bescherming persoonsgegevens* (2001)" further regulates the processing of personal data and the obligations associated with the processing of this data. First, it will be substantiated why this act is applicable to the PRSA method. Hereafter, conditions that must be met to legally process personal data are discussed. Subsequently, the rights of the data subject, "the person to whom personal data relate" (*Personal Data Protection Act (Unofficial translation)*, 2001), are explained. Finally this Act will be put into an international context, informing implementers of a PSRA system about how they can identify legal implications in their country.

Application

The "*Wet bescherming persoonsgegevens* (2001)" only applies to personal data, which is defined as "any information relating to an identified or identifiable natural person" (*Personal Data Protection Act (Unofficial translation)*, 2001). The aim of a PRSA system is to include publicly available personal information in risk analyses. In order to be useful for the risk analysis, that information should be related to the specific subject of that risk analysis. After all, if information is unrelated to the risk analysis subject, the analysis will be of no value to the decision making process. Therefore it can be concluded that a PSRA system, assuming that it works correctly, will process information related to an identified natural person.

Additionally, the act only applies "to the fully or partly automated processing of personal data, and the non-automated processing of personal data entered in a file or intended to be entered therein."

(*Personal Data Protection Act (Unofficial translation)*, 2001). Dividing this into two parts, the first part is about processing personal data. Although the verb *to process* has already been used several times in conjunction with personal data, it is essential that it matches with the definition of processing personal data within the *Wet bescherming persoonsgegevens* (2001), which is as follows:

> "processing of personal data" shall mean: any operation or any set of operations concerning personal data, including in any case the collection, recording, organisation, storage, updating or modification, retrieval, consultation, use, dissemination by means of transmission, distribution or making available in any other form, merging, linking, as well as blocking, erasure or destruction of data; (*Personal Data Protection Act (Unofficial translation)*, 2001)

Multiple verbs encountered in this definition are applicable to the system, such as "collection" and "retrieval". Therefor, it is evident that the system processes personal data. The second part, the process being automatically, is fairly obvious from the fact that the goal of a PSRA system is to support risk analyses with personal data from public sources automatically.

Conditions

In order to legally process personal data, certain conditions must be met. The first condition is that the data subject should have "unambiguously given his consent for the processing" (*Personal Data Protection Act (Unofficial translation)*, 2001). This means that prior to the process of utilising publicly available personal data within a risk analysis, the data subject should have agreed upon this processing. Related to this condition, it is obligated to inform the data subject for which purpose their personal data is processed. Only if the data subject agrees upon both the processing itself and the purpose of the processing, the personal data can legally be used. However, it is only allowed to use their data for the initial purpose they agreed upon.

Another obligation is that the personal data should not be kept any longer than necessary. This implies that, after the personal data has been utilised within the risk analysis, it should be deleted. Additionally, during the process and before the deletion of the personal data, it is obligated to "implement appropriate technical and organisational measures to secure personal data against loss or against any form of unlawful processing" (*Personal Data Protection Act (Unofficial translation)*, 2001). Which measures can be considered appropriate is outside the scope of this research, but implementers are recommended to obligate to this rule.

Before a system as proposed in this research is initiated, the responsible party must notify the Data Protection Commission that they will be automatically processing personal data. This way, the Data Protection Commission can maintain a database that supports them to enforce the law. The last condition that must be met requires the responsible party to not base any decision, which affects the data subject substantially, solely on personal data that has been automatically

processed. Therefore it is necessary to incorporate information in the decision making process that is not automatically processed, i.e. a report based on a conversation.

Rights

The person to whom the personal data relates has additional rights regarding the processing of his personal data in a system implemented by the PSRA process. First, each person has the right to request the responsible party to inform him as to whether personal data relating to him are being processed (*Personal Data Protection Act (Unofficial translation)*, 2001). The responsible party must then inform the data subject whether or not personal data related to him is being processed and, if so, provide a summary thereof.

In addition, the data subject is entitled to let the responsible party "correct, supplement, delete or block the said data in the event that it is factually inaccurate, incomplete or irrelevant to the purpose or purposes of the processing" (*Personal Data Protection Act (Unofficial translation)*, 2001) . These rights, however, will be of little impact to such a system since the time frame wherein the data is processed, before the personal data gets removed, is relatively small.

International context

Since the *Wet bescherming persoonsgegevens* (2001) is an implementation of the European directive 95/46/EG (Parliament & the Council of the European Union, 1995), it is quite likely that other European countries have similar acts in place. Therefor the PSRA process proposed in this research includes a step wherein research should be done into legal conditions. Another important aspect to keep in mind during implementation is that the act enforces that no personal data is send to a country outside the European Union, other than countries that guarantee an adequate level of protection. This is particularly important since the United States does not have an act to protect personal data at this time. As a result it is, for example, not allowed to make use of a cloud based tool whose storage is located in the United States.

3.5.3 *Summary*

In summary, the following issues should be taken into account when implementing a system such as the one proposed in this research:

1. The subject of the risk analysis should have unambiguously given his consent for the processing.
2. Along with the consent it is obligated to inform the data subject for which purpose his personal data is processed. It is only allowed to use their data for this purpose they agreed upon.
3. Personal data should not be kept any longer than necessary.
4. Appropriate technical and organisational measures to secure personal data against loss or against any form of unlawful processing should be implemented.

5. Before putting the system in operation the responsible party must notify the processing to the Data Protection Commission.

6. The responsible party must not base any decision, which affects the data subject substantially, solely on personal data that has been automatically processed.

7. The responsible party must be able to inform the data subject whether or not personal data related to him is being processed and, if so, provide a summary thereof.

8. The responsible party must be able to correct, supplement, delete or block the said data in the event that it is factually inaccurate, incomplete or irrelevant to the purpose or purposes of the processing when this is stressed by the subject.

9. No personal data may be sent to a country outside the European Union, other than countries that guarantee an adequate level of protection.

This list can be used as a guidance during the implementation of a system such as proposed in this research within the Netherlands. Since it is largely based on the European directive 95/46/EG (Parliament & the Council of the European Union, 1995) it might be useful within other legal jurisdictions. This list will also be used during the evaluation of the proof of concept in Chapter 6.

4

EXPLORATORY INTERVIEW RESULTS

4.1 FRAUD AND THE COMPANY

As already mentioned, the daily tasks of the experts are all related to fraud and thus fraud plays a role in each of the three companies. Expert II and III consider their customers (in the case of Expert II their *customers* are Dutch citizens) as the only group that could commit fraud. However, Expert I mentions their own employees as a group that could also commit fraud.

> Employees commiting fraud is fortunately a rarity, but we are not naive.

Although this research focuses on risk analyses of (prospective) customers, the research might also prove to be useful for risk analyses of employees. Section 4.4 provides an example given by Expert I where the proposed system in this research could also be used for the detection of fraud committed by the second group. For now the focus will be on the first group that could commit fraud, the (prospective) customers.

Due to the diverse sectors in which the three different companies operate, they also have to deal with various types of fraud. The company where Expert I is working is active in the e-tailing sector, hence they mainly have to deal with on-line purchases. In the case of business customers, fraud is committed by making a purchase on credit without ever actually paying the bill. In the case of individual customers the majority of the fraud cases are captured by external parties, namely in the case of phishing "the financial risk is not for us, but for the bank". And in the case of stolen deliveries "the risk is for [...] the postal delivery company". A third type of fraud committed by individual customers is at the company's own risk and is related to credit card transactions, because it is possible to perform a chargeback with a creditcard even when the order has already been delivered. At the company that is active as a credit provider each application for a credit could potentially be fraudulent. Credit applicants falsify identification documents, payslips and bank statements. Additionally, credit applicants also deliberately lie about their age, job, household, etc. to influence their financial profile. Finally, the government organization Inspectie SZW probably has to deal with the biggest diversity of types of fraud of the three companies. Within the social affairs and employment sector there are a lot of ways in which fraud can be committed. Because there are so many the organization works with a project-based approach where they pick up some types of fraud per project, and not all at the same time.

These findings show that risks in terms of fraud play a role in some very diverse sectors, besides it is not impossible to imagine that fraud plays a major role in various other sectors as well. And since it probably plays a role in so many places an improvement in

the mitigation of these fraud related risks, which is the aim of this research, may also be of value in many places.

4.2 MANUAL MEASURES

Although manual measures that mitigate fraud related risks are not directly relevant to this research it does provide context and might provide future possibilities for the automation of these manual measures. At the top 3 e-tailing company many cases of fraud are prevented by not allowing individual customers to buy on credit. In addition, they also purchase credit assessments about both business and individual customers from external parties. They do this to get informed about the solvency of their customers.

> For business to businnes it is actually quite simple, we purchase credit assessments from several companies. [...] What we do, of course, — just like any other business – is buying credit assessments [for individual customers].

At the credit provider they have an "acceptance team of roughly 20 persons" in place that manually checks each credit application for indications of fraud. They are informed about every aspect of the identification documents, payslips and bank statements such as the security features, font, spacing, lay-out and coporate colors of the company, etc. As previously mentioned the Inspectie SZW works with a project-based approach, therefore they do not have general manual measures in place.

At all three companies manual investigations are carried out into possible fraud cases. These investigations are started when there are indications of some sort that a customer attempts to commit fraud. In some occasions, at the Inspectie SZW and the credit provider, these indications may also originate from an automated system, which will be discussed in Section 4.3. During these manual investigations all kinds of techniques are utilized including, but not limited to: manual lookups in their own systems, manual lookups in external systems, contact with other companies and contact with the customer themselves.

Some manual measures such as acquiring a credit assessment of a (prospective) customer in the case of the e-tailer company can be automated. This has actually already been automated by the credit provider company, every applicant's solvency is automatically checked by their system. Although this is a great example of a manual measure that can be easily automated it is indicated by expert III that not everything can be automated. He describes a certain *feeling* that the acceptance team and himself have when they get to see a credit application, a feeling based on experience in the field. According to him it is nearly impossible to automate this feeling with the use of a system.

> Some employees have been here for year, or decades, and they have developed such a unique feeling; they just pick them up. And then it is not even necessary for me to look at it, I just now it is a fraud. [...] And I think it is hard to achieve the same result with an automated system.

4.3 AUTOMATED SYSTEMS

Another purpose of the exploratory interviews was to identify automated systems related to risk analyses, fraud detection and fraud prevention, as not everything that is done in business is externally published. At the moment, the top 3 e-tailer company has no such automated system in place at all. Expert I does see potential in an automated system, specifically in a system that assigns risk points to customers when they, for example, change their delivery address. When a customer has a relative high amount of points the order is investigated for fraud.

> I would like to developed software for that, which makes it possible to assign risk points and that these risk points — in time — erode away.

The company that provides credits does have an automated system in place. This system automatically runs several tests on a credit application in order to estimate the risk for the purpose of detecting and preventing fraud. It retrieves the credit applicant's solvency from the central register in the Netherlands(Bureau Krediet Registratie), it checks if the identification document is registered in the central register of stolen and lost identification documents, it checks their internal systems for previously registered fraudulent activities and it checks the central system for fraudulent activities in the Netherlands, which is called Externe Verwijzings Applicatie (External Reference Application; EVA). These tests do not only return a positive or negative result, but also search for other indicators of a fraudulent credit application. As an example the expert explains that the central register in the Netherlands also returns the address of a credit applicant, if this differs significantly from the address of the credit application the system also flags it as possibly fraudulent.

> Address data is also returned. So when someone has recently applied for a credit in Utrecht, and tells us out of nowhere that he lives in Groningen the system tells us to watch this person, to double check if it is correct.

In addition the system also features filter functionality that makes it possible to automatically flag a credit application as possibly fraudulent based on a set filter. An example of such a filter is that all credit applications of people who say that they work at a given company are automatically flagged as possibly fraudulent.

The government organization Inspectie SZW seems to have the most advanced automated system, named Risico Analyse Omgeving (Risk Analysis Environment, RAO), of the three companies in place. This system extracts data from a large number of systems including, for example, systems of tax authorities and municipalities. The system then utilizes this information to carry out risk analyses and presents a list of, for example, potential fraudsters. In order to do this, the RAO system makes use of so-called risk-indicators that are defined by analysts such as expert II. Each indicator has a value that counts towards a score, and persons that have a relatively high score

are presented to the inspectors. The persons on this list are then further investigated. The persons that do not stand out based on their score are not accessible by the analysts, thus their identity remains unknown. For a more extensive description of the system the reader is referred to Appendix C.

These findings show that automated systems for risk analyses in fraudulent environments are already in place in some companies. Although the top 3 e-tailer company does not have an automated system in place at the moment, expert I does see potential in such a system. This research focuses on extending an existing risk analysis system and hence this research is less applicable to that specific case. However, the other two systems identified offer opportunities to be extended with personal data from public sources. The system at the credit provider already flags credit application when, for example, addresses retrieved from the connected systems differ significantly from the address on the application form. The same could be achieved by comparing personal data from public sources with personal data on the application form. The RAO system, which is in place at the government organization Inspectie SZW, already extracts data from multiple sources. A public source such as Facebook could be added as an additional source and data extracted from it can be used in the same manner as data extracted from other sources. Whether the experts see potential in these extensions is discussed in Section 4.5. The next section will first concentrate on the current manual use of personal data from public sources within the companies.

4.4 MANUAL USE OF PERSONAL DATA FROM PUBLIC SOURCES

As discussed in Section 4.2 all experts state that they carry out manual investigations once a possibly fraudulent case has been indicated. This indication can originate from both a manual measure as well as from an automated system. All three experts state that they utilize personal data from public sources such as Facebook during this investigation, but not at all on other moments. This means that an automated system or a manual measure has already flagged a subject as suspicious when personal data from public sources comes into the picture. Expert I provides a real-life example of a customer that redeemed an abnormal amount of gift vouchers, which caused the company to start an investigation. Using the friends list on the customer's profile, they identified a connection between the customer and one of the employees of the company. Further investigation lead to the conclusion that these two collaboratively committed fraud.

> If someone redeems 30 — or 40 or 50 or 60 — gift vouchers in a week than bells start ringing. [...] And then, it helps when people list their connections publicly so we can see whether there are employees among them. [...] And this has happened one or two times in the past 10 years that we were able to connect [a fraudster] with an employee.

The current use of this type of data is fundamentally different than the system proposed in this research, which aims to flag a subject as suspicious (partly) based on personal data from public sources.

However, from the fact that the experts indeed use this kind of data to fulfill their daily tasks it is evident that they at least consider the data usable for their work. The question is whether they also consider the data usable enough for the next step, flagging subjects as suspicious based on this data. This question will be answered in the following section.

All experts indicate that they never solely use personal data from public sources to decide whether it is a case of fraud. They combine it with other sources or use it only as a guidance for their research. This finding nicely aligns with the way in which the system in this research is proposed: as an extension to already existing systems instead of an independent system that only utilises personal data from public sources.

4.5 AUTOMATED USE OF PERSONAL DATA FROM PUBLIC SOURCES

As none of the companies already used personal data from public sources in an automated manner, the question arose why they do not make use of it yet and whether they do see potential in such a system at all. The experts gave multiple reasons why they do not see potential in a risk analysis system that utilizes personal data from public sources.

First, expert I states that personal data on public sources such as Facebook is "unstructured and diffuse". This problem is also one of the main problems that this research tries to solve. The literature review in Section 3 discusses techniques how these two problems can be solved. The problem related to the data being unstructured is solved with techniques from the web information extraction research area. The techniques allow to retrieve structured personal data from the public sources with the use of wrappers. The problem related to the data being diffuse is solved by both the adoption of techniques from the business intelligence area as well as techniques from the entity matching area. The techniques from the business intelligence area allow to gather data from the different sources on which the data is spread and load into a single data warehouse. Hereafter, entity matching makes it possible to only keep the data related to the subject of the risk analysis.

A second reason given by expert I, is that the limited number of cases wherein personal data from public sources can contribute to a fraud investigation does not outweigh the implementation effort of such a system. Especially expert I indicates that they only use these sources in exceptional cases. Expert II and III state that they do use these sources at least on some regular base to obtain background information about the subject. This may indicate that the system could be more useful at their companies. Theoretically, when the system is able to provide a substantial increase in the detection and prevention of fraud it should return the investment quickly.

Another reason, confirmed by all experts, is that people can deliberately create or change a profile on a public source to influence the outcome of risk analyses. Expert I mentions that it is no problem at all to commit fraud this way.

> But someone that really wants to commit fraud, that is no problem at all. You can build a fake Facebook profile if you know how it works, if you know how the checks for fraud take place than you can influence all of it.

This is indeed one of the major liabilities of the proposed system as none of the data on these sources is verified. This is in contrast to the data extracted from systems of, for example, the tax authorities which is extensively checked by their composer. Expert II does provide a possible solution for this problem. He states that if they are ever going to make use of social media as a source in their automated system, they would assign the indicators that are based on data from public sources a relatively low value. However, this would also decrease the influence of the personal data on these sources and that might diminish the usefulness of the extension.

Directly related to the previous reason, all of the experts state in at least one manner that the data on public sources is unreliable, even when a profile is not deliberately created or adjusted to influence the outcome of risk analyses. Expert II indicates as an example that a post about the purchase of a new Porsche might "just be a bluff". Again, nothing on these sources is verified. Additionally, according to expert III the information on these sources is also often "not up-to-date". This may be caused because some people simply do not update their information or might have switched to another social network. Apart from the fact that unreliable data is not a good base for a decision, expert III states that it would also cause a lot of unnecessary work. This would happen when the personal data on these sources does not match with the personal data in the other sources, for example the current employer. An unnecessary fraud investigation would then be started solely because the subject did not update his current employer on the public source.

> If I look at myself, when I create a [social media] profile, I do not update it each month or half year. I can not remember that I changed my employer somewhere, so on some of my profiles there are probably still names of employers from about 8 to 9 year ago. So when I would apply for a credit you can not use the information to determine that I work for company A and say I work for company B. Then a whole verification process would be started, which is not necessary at all.

This problem can be solved in the same manner as the problem with the deliberately created or changed profiles, by assigning a low value to indicators based on data from public sources, but once again that might affect the usefulness of such a system.

Finally, some minor specific reasons are indicated by the expert. Expert II indicates that there are a lot of systems spread through the Netherlands which are not yet connected to their system and have a higher priority to be implemented.

> We are willing to undergo that [including public sources], we are willing to take a look into that, but there are so

> many other systems which are basically mandatory for us. Those Basic Registrations, and we prefer to choose those instead of getting involved with social media.

The reason why the other sources have a higher priority is directly related to the low reliability of the public sources, the other systems are considered to be a lot more reliable. Expert III indicates that such a system would slow down their process of credit application approval, one of their unique selling points and crucial in their type of business. However, with enough performance it should not take much longer to compare personal data on public sources with the application form data then comparing this with data from a central register. Besides, it would certainly not take longer than manually contacting a bank to verify information. Expert II agrees with the feeling that the amount of information that people make publicly available decreases, possibly due to more privacy-oriented default settings of sites such as Facebook. If this trend continues the usefulness of the system would further decrease.

In summary, the three experts indicate quite a few reasons why they do not use personal data from public sources in an automated matter and also do not see a lot of potential in such a system. Some of the problems indicated can be solved or overcome, such as the data being unstructured and diffuse, the limited use case of the system and the delay the system might cause. Other problems, such as the unreliability of the personal data on public sources, the possibility that people might deliberately create or change their personal data on public source and the amount of personal data publicly available, fundamentally affect the usefulness of the system. These reasons make it apparent that it should be investigated appropriately whether this system adds something to the current fraud detection and prevention measures before implementing it.

4.6 ETHICAL ISSUES

Part of the exploratory interviews was to gauge the opinion of the experts related to ethical issues. In general, none of the experts see major ethical issues when using personal data from public sources. Expert I indicates that on one hand using Facebook for fraud detection and prevention is not where it was primarily intended for. But on the other hand he notes that it could also serve for the protection of their customers, for example when they have become a victim of phishing. In addition he notes that when the information is explicitly made public it can be used for other purposes as well.

> I do think that Facebook is part of the personal space, it is not where Facebook is primarily intended for. [...] But on the other hand, if it is publicly available information, you are allowed to use it.

Expert II agrees with the latter statement and sees no ethical issues at all. According to him whenever a user makes use of a service he should now how to use that service properly. Expert III says that it is too short sighted to say that it is entirely the responsibility of

the user, as some services are or were not privacy-oriented in their default settings. However, according to him credit providers already have a shortage of options for the prevention and detection of fraud so he uses whatever exists as long as it legally allowed.

> On the other side, there is not so much left that a credit provider can use to test the truth. [...] I can imagine that [people have objections], but on the other side when I can use it to prevent fraud, I will do that with all the love and pleasure.

Generally speaking, the experts do give the impression that they understand that people might have ethical objections to the use of personal data from public sources for risk analyses. However, all the experts nevertheless make use of the opportunity once it arises. They give the impression that other interests, such as their own fraud detection and prevention or those of a victim of phishing, outweigh these ethical objections. These are not entirely unexpected findings as the experts believe in the importance of their work and will use whatever they can as long as it is legal. A future study among the customers regarding these risk analyses will most likely yield entirely different results.

5 PUBLIC SOURCES FOR RISK ANALYSES

As an answer to the main research question, stated in Section 1.2, this chapter introduces two artifacts related to the implementation of a system that extends current risk analysis systems with personal data from public sources. The first artifact is a process that will guide implementers of the system during the implementation thereof. The phases and steps in this process are based upon the artifacts identified in Section 3, such as the research areas themselves, their methods and techniques and the legal literature review. The second artifact is an architecture, this architecture is intended as a high-level reference architecture of the system. This high-level reference architecture is also based upon the artifacts identified in Section 3, including the architectures from the Business Intelligence research area and the methods and techniques from the Web Information Extraction and Entity Matching research area. These two artifacts are called the PSRA Process and the PSRA Architecture, respectively.

First, the PSRA Process will be presented in the next section. The main phases of this process will be further elaborated in separate subsections, along with the steps that these phases contain. Hereafter, the PSRA Architecture will be presented. The internal components of the architecture will also be addressed in separate subsections. Because the PSRA Process guides the implementation of a system of which the architecture is based on the PSRA Architecture, it is evident that multiple relations between the two artifacts exist. These relations will be addressed in the sections whenever they are relevant.

5.1 PSRA PROCESS

As already mentioned, the purpose of the PSRA Process is to guide implementers during the implementation of a system that extends current risk analysis systems with personal data from public sources. The process itself consists of six sequential phases and each phase contains multiple steps, these steps should be executed in a sequential order as well. The phases of the process are visually depicted in Figure 5.1.

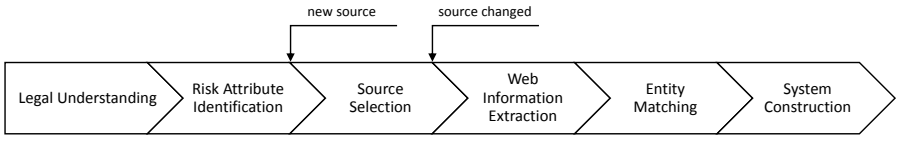

Figure 5.1.: Phases of the PSRA Process

The first phase is Legal Understanding and its aim is to get an overview of the local legal issues that should be taken into account during the development and the operation of the system. The second phase, Attribute Identification, is focused on the identification of the most valuable risk attributes that should be extracted from the public sources. Source Selection is the third phase, and its goal is to create

an ordered list with the most valuable public web sources, based on the extent to which the earlier identified risk attributes are present on these sources. The fourth phase is Web Information Extraction, this phase is based on the Web Information Extraction research area and its aim is to guide the development of wrappers specific for the sources identified in the previous phase. Entity matching is the second to last phase in the process. It is based on the Entity Matching research area and its goal is to guide the development of the source-specific matchers that will decide which extracted profile refers to the same real-world person as the subject of the risk analysis. Finally, the sixth phase focuses on the integration of the various components and the existing risk analysis system as well as the testing and deployment of the final system.

In addition to the phases, two events are also included in Figure 5.1. The first event that might take place, during the execution of the process or after the system has been deployed, is that a new source is discovered. This could be a source that already existed, but was not yet discovered. Or it could be a newly introduced source. When this even takes place, the process should be, partly, restarted from the Source Selection phase.

The second event that might take place is that a source might change. When a source changes, and a wrapper has already been developed for this source, there is a high probability that the existing wrapper does not function correctly anymore. Therefore, the process should be, partly, restarted from the Web Information Extraction phase.

5.1.1 *Legal Understanding*

The first phase in the PSRA process is Legal Understanding. The aim of this phase is to get an overview of the legal issues that should be taken into account during the development and the operation of the system. Especially when implementing a system that processes personal data, such as the one proposed in this research, it is important to be informed about the legal restrictions and obligations in advance. This is because these legal restrictions and obligations can have a major impact on the implementation process and the implemented system itself. When these restrictions and obligations are known beforehand, the process and the implementation itself can be adjusted to meet these requirements and, potentially, save a lot of time and money. After all, when these requirements have to be met afterwards it takes more work than when it would have been part of the initial development. Let alone the possible fines that the responsible party might receive because the legal obligations are not met or the negative publicity that the responsible party might receive.

This phase is part of the PSRA Process because the it was discovered in the legal part of the literature review in Section 3.5 that legal restrictions and obligations differ between countries. There are countries that are very restrictive in terms of privacy, and countries that are not so restrictive in terms of privacy. Therefore, extensive research should be done in this area before each implementation of a system such as proposed in this research. Legal understanding is the first

phase because it could potentially affect the rest of the process, as already discussed above. It is recommended to execute the steps in this phase together with a legal consultant.

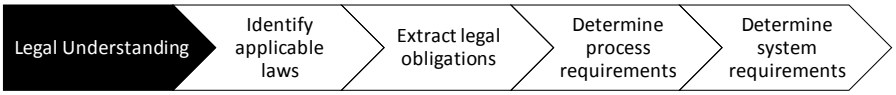

Figure 5.2.: The Legal Understanding phase of the PSRA Process along with its steps

As can been seen in Figure 5.2, this phase consists of a fourfold of steps. The first step is to identify the laws that are applicable at the location where the system will be implemented. Hereafter, in the second step, legal obligations will be extracted from these laws. In the third step additional requirements for the implementation process are determined based upon the extracted legal obligations that affect the overall process. In the fourth, and final, step requirements for the system itself are determined based upon the extracted legal obligations that affect the system.

IDENTIFY APPLICABLE LAWS During the legal study in Section 3.5 it was discovered that multiple laws can influence the implementation of a system such as proposed in this research. Therefore, in order to achieve legal understanding at the local level, it is important to firstly identify the laws that are applicable in that specific area. The easiest way to identify these laws is to acquire legal advice from a legal, local, consultant. This consultant should be able to inform the implementers, based on a description of the system, which laws are applicable to the implementation.

Besides acquiring legal advice from a legal consultant there are also other ways to get informed about the applicable laws. As discussed in Section 3.5.1 the European Commission for Democracy through Law maintains a database which systematically indexes the constitutions of numerous countries. This database is publicly available on `www.codices.coe.int` and allows anyone to effectively search for parts of the constitution. By using the indexes displayed in Table 3.2, implementers of the PRSA system can search for relevant parts of the constitution applicable in their country. Often these relevant parts contain references to other, more specific, applicable laws such as in the Constitution of the Netherlands (*Grondwet voor het Koninkrijk der Nederlanden*, 2008). However, not all applicable laws are likely to be refered to in the Constitution. So, there should also be sought in other ways.

Whenever the country wherein the system is implemented is part of the European Union, it is likely that there is also a more specific law applicable that protects the subject's personal data. There is the European directive 95/46/EG (Parliament & the Council of the European Union, 1995) that serves as a guideline for further implementations of a personal data protection act. Implementers can search for implementations of this directive in their own legal field. Independent from the way in which the

applicable laws they are identified, the final deliverable of this step is to compose them into a list.

EXTRACT LEGAL OBLIGATIONS It was found, during the legal study presented in Section 3.5, that the applicable laws contain obligations that should be met during the implementation. Therefore, after the applicable laws have been identified, the next step is to go through these laws thoroughly. During this process all obligations, contained within applicable parts of these laws, should be extracted and written down. Again, advice from a legal consultant can ease this process, or even eliminate the need for this process at all, as he should be able to provide this list of obligations. In addition, it is also possible to search for legal documents that already summarize the applicable laws. However, those might not fully cover all contents of the specific law, a risk that has to be taken into account. The deliverable of this step is a complete list of obligations extracted from the applicable laws.

As an example of a legal obligation extracted from an applicable law, the following is taken from the literature review in Section 3.5.3:

> Before putting the system in operation the responsible party must notify the processing to the Data Protection Commission.

After the *Wet bescherming persoonsgegevens* (2001) had been identified as an applicable law, one of the sections within the law obligates parties that are responsible for the processing of personal data to notify the Data Protection Commission before the processing is started. From this section, the above stated obligation was extracted and written down. When this is done for each applicable law, the deliverable of this step, a list of legal obligations that should be met by the responsible party, is completed.

DETERMINE PROCESS REQUIREMENTS As mentioned above, the researcher noted that the legal restrictions and obligations identified during the legal study can affect the implementation process of the system. Therefore, this step focuses on determining the additional requirements of the PSRA Process. First, it is essential to examine which obligations from the list composed in the previous step affect the process. All obligations on the list are evaluated one-by-one whether they affect the process, when this is the case the obligation is added to a new list. The result is a list with merely obligations that affect the process, which is a subset of the list composed in the previous step. Additionally, it should be examined for each obligation what is required to be changed in the process in order to meet this obligation. The final result is a list of additional requirements of the PSRA Process that, when implemented, ensures that the extended process fulfills this part of the obligations.

As an example, this step is demonstrated with the example obligation chosen in the previous step. This obligation has been

taken from the complete list of obligations as listed in Section 3.5.3. This obligation affects the process because somewhere before the system is deployed and initiated, the Data Protection Commission should be informed by the responsible party about the future processing. Therefore, it is evident that this obligation is added to the list of obligations that affect the process. In order to fulfill this obligation, a requirement specific to this obligation is formulated as follows:

> The PSRA Process should contain a step, before the system is able to process personal data, wherein the Data Protection Commission is notified of the future processing.

This should be repeated for all other obligations on the list with obligations that affect the process to produce the final result of this step, a list with additional requirements of the PSRA Process.

DETERMINE SYSTEM REQUIREMENTS As the previous step was concerned with determining the additional process requirements, this step is concerned with determining the additional system requirements. During the legal part of the literature review in Section 3.5 it was also noted that some of the restrictions and obligations affect the system itself. Therefore, this step is included within the process. Determining the system requirements is done in the same manner as determining the process requirements. First, the list with obligations is examined one-by-one to identify obligations that affect the system. These form a new list with obligations that affect the system and, again, this is a subset of the list composed in the Extract Legal Obligations step. Note that this subset might overlap with the subset of obligations that affect the process, as a single obligation can both affect the process as well as the system. Based on each of these obligations, additional requirements for the system are made in order to fulfill this part of the obligations. This results in the final deliverable of this step, a list with additional requirements for the system.

As an example, the following obligation has been taken from the complete list of obligations as listed in Section 3.5.3:

> Personal data should not be kept any longer than necessary.

This specific obligation affects the system because technical measures should be taken in order to make this possible. Therefore, this obligation is added to the list of obligations that affect the system. A requirement specific to this obligation is formulated in order to fulfill this obligation:

> Personal data should be deleted immediately after it is decided that this data is not related to a subject of the risk analysis. Personal data related to a subject of the risk analysis should be deleted immediately after the decision is made for which the data was extracted.

This should be repeated for all other obligations, on the list with obligations that affect the system, to produce the final result of this step, a list with additional requirements of the system.

5.1.2 *Attribute Identification*

Attribute Identification is the second phase in the PSRA Process. This phase is focused on the identification of the most valuable risk attributes. These risk attributes are the attributes that will be used by the analysis tools to decide whether the subject of the risk analysis is a potential fraud. Examples of risk attributes are: the subject's current employer, marital status, age, etcetera. Since several subsequent steps, as well as the final result of the risk analysis, depend on the identified attributes, this is a crucial part of the process. The final goal of this phase is to produce a prioritized list of risk attributes specific for the domain wherein the system is implemented. In the next phase, Source Selection, this prioritized list of risk attributes will be used to select the most useful sources for the system.

This phase is included in the PSRA Process because it was found, as discussed in the interview results in Section 4, that the type of fraud differs among the various domains. When the type of fraud differs, the way in which the fraud can be detected and prevented also differs. This implies that the attributes, which are considered important to identify the fraud, should be individually identified for each implementation of the system. A credit provider, for example, would value the attribute *current employer* higher than that an e-tailer would value the same attribute. Therefore, the identification of the risk attributes is an important part of the PSRA Process.

Figure 5.3.: The Attribute Identification phase of the PSRA Process along with its steps

The Attribute Identification phase consists of a fourfold of sub steps. The first step is to select experts of the specific domain wherein the system is implemented. The next step is to conduct an interview with each of these experts in order to extract their knowledge and experience in this specific domain. Hereafter, the results of the interviews are used to extract attributes from. In the final step these extracted attributes are prioritized. The entire phase is visualized in Figure 5.3.

SELECT DOMAIN EXPERTS The first step to identify the most valuable risk attributes, is to identify and select domain experts. As discussed above, it was found during the exploratory expert interviews in Section 4 that the types of fraud and the way in which fraud can be identified differs among the various domains. A method to acquire this specific knowledge of the domain at-hand, is to conduct interviews with experts of that domain. This step focuses on selecting the experts for those interviews.

When selecting domain experts, it is recommended to search for people who deal with fraud within that domain on a daily basis. The experts, which are most likely to possess useful knowledge, are the experts that currently carry out manual fraud investigations. These should be able to describe attributes that they use to detect and prevent fraud. Other potentially interesting experts are the ones that possess knowledge about the automatic systems, related to fraud detection and prevention, that are already in use. They should be able to inform the implementers about the attributes these systems use for the detection and prevention of fraud.

The following applies to the amount of domain experts for interviews: the more the better. The more interviews conducted with different domain experts, the greater is the chance that all of the most important attributes are identified. Additionally, more data will be available for the prioritization of the attributes later on. The result of this step should be a list of domain experts that can be used in the next step, Conduct Interviews.

CONDUCT INTERVIEWS After the domain experts have been selected, the next step is to conduct interviews with these experts. Before the interviews are conducted, an interview protocol should be created to ensure the quality of the interview results. An interview protocol makes sure that all points that should be addressed are actually addressed during the interview. Appendix B is an example interview protocol used during this research. A semi-structured interview is recommended since the open concept of a semi-structured interview allows points to be addressed that the interviewer had not thought of beforehand.

The focus of the interview should be on the identification of potential valuable risk attributes. Several possibilities exist to identify these and this also varies from case to case. When the expert already uses attributes from public sources during a manual fraud investigation, he can be asked for the attributes he generally uses. If he does not know which attributes he generally uses, he can be asked for specific examples of previous fraud investigation and the attributes that proved to be valuable during that investigation. If public sources are not used at all during manual investigation, one could ask for examples of other sources they use during their investigations. Personal data in these other sources may as well exist in public sources, or are in any case comparable to personal data on public sources. This is certainly not a complete list of possible ways to identify valuable risk attributes, it should be noted that this step requires some creativity and adaptability in order to be taken.

Finally, each interview should result in a document that contains the results from the interview. Whether this document is a full transcription of the interview or a good summary does not matter. As long as the key points related to potential valuable risk attributes are all included. The set of documents with the interview results are the final deliverable of this step.

EXTRACT ATTRIBUTES After the interviews have been conducted, and the results thereof are processed into documents, the next step is to extract the actual attributes from these documents. Each document produced in the previous step should be examined one-by-one and should be searched for potential valuable risk attributes. Whether they can be extracted directly, or they should be deducted from the document, depends on the question to which it was an answer. If the question was directly about the attributes they already used from public sources, they can be extracted directly. In all other cases more creativity and adaptability is again required. For example, when an expert indicates that they check whether the address provided by the customer matches their address in the central register, the address of the customer on public sources could be used in the same manner. Each time when a new attribute has been extracted from a document, it should be written down. Additionally, a short summary of the reason why this attribute is considered important should be written down as well for documentation purposes. This could be, for example, a description of an investigation wherein this attribute proved to be useful, or the purpose for which the expert indicated he uses this attribute. Whenever a specific attribute has already been extracted before, a note of this new occurrence should be added to the original entry as this may be useful during the next step. If the new reason differs from the original reason, the new reason should be added to the original entry as well. This way the deliverable of this step, a list with extracted attributes, an accompanying reason and the number of occurrences, is composed.

PRIORITIZE ATTRIBUTES The last step in the identification of the risk attributes is to prioritize the list, composed in the previous step, based on the perceived importance of the risk attributes. The idea behind this is that the experts indicated, during the exploratory interviews from Section 4, that not all potential risk attributes are considered to be equally important. As an example, one can imagine that when there is a mismatch between the age that a customer specifies and the one extracted from a public source, this is considered to be more important than a mismatch in the primary school they attended.

There are several possibilities to prioritize the list. The easiest way would be to count the amount of times that each risk attribute was identified during the interviews. When multiple experts in the same domain are interviewed thoroughly, it is very likely that from these interviews the same attribute will be extracted multiple times. This way, all that is needed is to keep track of the occurrences during the Extract Attributes step. These occurrences are then used as the importance value for the risk attributes.

Another way would be to manually assign the importance values of the risk attributes. This can be done based on the importance that is perceived from the interviews. Experts, for example, might have explicitly stated that certain attributes are

Attribute	Importance value
Age	2
Address	2
Employer	3
Primary school	1

Table 5.1.: Example deliverable of the Attribute Identification phase

more important than others. But the number of occurrences could also be taken into account in manually assigning the importance values. In addition, some rules can be set to guide the assignment, for instance:

1. Attributes that have the same importance values are considered approximately equally important.
2. Attributes with a lower importance value are considered less important than attributes with a higher importance value.
3. When attribute A has an importance value twice as high as attribute B, attribute A is considered approximately twice as important as attribute B.

The result, which is also the final deliverable of the Attribute Identification phase, is a list with risk attributes prioritized by their importance value. A simple example can be seen in Table 5.1. This list will form the base for the next phase, the Source Selection phase.

5.1.3 *Source Selection*

After the most valuable risk attributes have been identified in the previous phase, the goal of the third phase is to select the most valuable public sources. The selected public sources will be utilized by the system to extract the attributes from. Since the integration of a source is a time, and thus money, consuming operation, it is important that the most valuable public sources are integrated first. After all, it would be a waste if a lot of resources are spent on the integration of a source from which only two relatively unimportant attributes can be extracted. Especially when another public source, which contains multiple relatively important attributes, is not yet included in the system. At the end of this phase, a prioritized list should have been composed of public sources available within the specific location.

The Source Selection phase is part of the PSRA Process because it differs among the domains, as well as among the implementation locations, which sources are the most valuable. As previously discussed, different domains resulted in a different set of valuable risk attributes. Since these risk attributes should be extracted from the public sources, the most valuable public sources might also be different between two distinct domains. The specific location also matters because public sources can also be location specific. On the one hand, a source could only be available within the specific location, for example a social network aimed at a specific country. On the other hand, some very wide-spread social network, for example, might not yet be

available in the specific location. Because both the domain and the location co-determine the most valuable sources, this should be examined both before each implementation of the system and therefore it is an essential part of the PSRA Process.

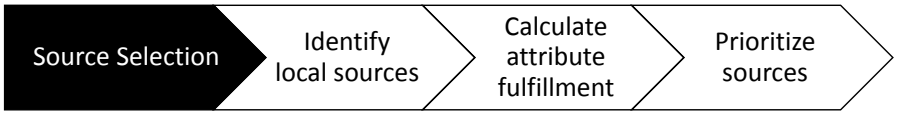

Figure 5.4.: The Source Selection phase of the PSRA Process along with its steps

This third phase, depicted in Figure 5.4, contains three steps that guide the selection of the public sources that will be included in the system. The first step is to identify the public sources that are available within the specific location. After the local public sources have been identified, a score is calculated for each of the public source that indicates the extent in which they contain the attributes identified in the previous phase. In the final step of this phase, a prioritized list of the sources is composed, partly based on the attribute fulfilment score.

IDENTIFY LOCAL SOURCES In order to determine which public sources are the most valuable for the implementation at-hand, it is necessary to firstly identify as much, locally available, public sources as possible. An easy way to make a start is to write down the public sources that are used by the implementation team itself. Additionally, some sources might have already been mentioned by experts during the interviews in the previous phase. Another possibility would be to ask existing customers what kind of public sources, such as social network sites, they use. This way, as a bonus, the perceived popularity of that social network site under the target audience is also acquired, something that can be useful for the last step of this phase. Of course, more ways can be devised to identify local sources. However, the final goal of this step is to compose a list of public sources available at the implementation location. An example list can be seen in Table 5.2

Source (Netherlands)
eBay
Facebook
Google+
LinkedIn
Marktplaats
Twitter

Table 5.2.: Example list of locally available public sources in the Netherlands (not a complete list)

CALCULATE ATTRIBUTE FULFILLMENT Once as much public sources as possible have been identified, the next step is to calculate a score for each of them that indicates the extent in which the source contains the attributes identified in the previous phase.

For this, the list of attributes — along with their importance value — from the Attribute Identification phase is used together with the list of locally identified sources. For each public source from the list, a calculation is done based on *all* attributes identified. This calculation results in a score, the attribute fulfilment score, for each identified public source. This score should resemble to what extent that particular source fulfils in the need of the most valuable attributes. A variety of ways can be devised to perform this calculation, as longs as it results in a higher score for a source that contains more valuable attributes than another.

As an example the prioritized list of attributes produced in the preceding phase, depicted in Table 5.1, is used together with the list of locally available public sources produced in the preceding step. For each source it is checked which of the identified attributes can be extracted from that specific source. When an attribute can be extracted from the source, the source is awarded the importance value of that attribute. The sum of all the awarded importance values is used to determine the attribute fulfillment score. When all attributes could be selected from a particular source, that source would have an attribute score of 100%. Therefore, the total of awarded importance values is divided by the maximum score possible, this results in the attribute fulfilment score.

Source	Age (2)	Address (2)	Employer (3)	Primary School (1)	Total	Attribute fulfillment
eBay	-	2	-	-	2	25%
Facebook	2	2	3	1	8	100%
Google+	2	2	3	1	8	100%
LinkedIn	2	2	3	1	8	100%
Marktplaats	-	2	-	-	2	25%
Twitter	-	2	-	-	2	25%

Table 5.3.: Example matrix of the Calculate attribute fulfillment step

A convenient way to calculate the attribute fulfilment score is to create a matrix wherein the sources are listed vertically and the attributes are listed horizontally. On each intersection, the attribute importance value is entered whenever this attribute can be extracted from the source. At the end of the row the points are summed up, and divided by the maximum score possible. The matrix for the examples is depicted in Table 5.3.

PRIORITIZE SOURCES The final step of this phase is to prioritize the list, composed in the previous step, based on the perceived importance. This step is based on discovery, made during the exploratory interviews in Section 4, that the experts do not consider all sources equally important. Just like the prioritization of the attributes in the preceding phase, the prioritization in this phase can be done in several ways. One way is to use the

attribute fulfillment score of the sources to mutually rank them. This implies that the sources that have the highest attribute fulfillment score, will be the first on the list, and thus implemented first. However, multiple sources could have the same attribute fulfillment score or some sources might be barely used in practice. Therefore it is recommended to use the attribute fulfillment score only as a guidance during the prioritization of the sources.

Rank	Source (Netherlands)
1	Facebook
2	LinkedIn
3	Twitter
4	GooglePlus
5	Marktplaats
6	eBay

Table 5.4.: Example prioritized list of locally available public sources in the Netherlands (not a complete list)

It was previously mentioned, during the Identify local sources step, that the perceived popularity of a source would be bonus later on. Of course, it is also possible to use the market shares of the different sources if these are known. This can be extremely helpful during the final prioritization of the sources. When one knows that a particular source is only used by a small portion of the population, it is — in most occasions — not wise to implement this source first. Maybe not even when it fulfills a large part of the attributes. Apart from the perceived importance, other characteristics can be taken into account as well. When the list is prioritized, independent of the way in which it was done, the final goal of the Source Selection phase is reached. The prioritized list, resulting from this step, can be seen in Table 5.4

5.1.4 *Web Information Extraction*

The fourth phase in the PSRA Process has its origins in the Web Information Extraction area, which was discussed in Section 3.2. The aim of this phase is to guide the development of the wrappers, specific for the sources identified and prioritized in the previous phase. The wrappers developed during this phase will extract the data from those sources, after which the matchers developed during the next phase will utilize this data to determine which profile belongs to the subject of the risk analysis. As found during the literature review, a wrapper for a specific source should consist of two separate wrappers, a search results wrapper and a profile wrapper. The search result wrapper should wrap the search engine of the public source, which enables the system to make a preliminary selection of potentially relevant profiles. Once a list of potentially relevant profiles is obtained, the profile wrapper should extract the attributes from each of those profiles. This set of wrappers is developed for several of the identified sources, and these sets are combined into one module.

This module, called the wrapper module, is the final deliverable of this phase.

This phase, the Web Information Extraction phase, is included in the process because the sources vary among the different domains and implementation locations. Therefore, wrappers have to be developed for sources that might not have been valuable during earlier implementations in different domains and locations. Additionally, the wrappers developed during this phase are essential for the functioning of the entire system. These wrappers extract the personal data from the public sources, and without this data the rest of the system would not function. The matchers, developed in the next phase, utilize this data to determine which profile belongs to the the subject of the risk analysis. The analysis tools that will be used to perform the actual analysis, which are outside the scope of this research, use this data to determine whether the subject of the analysis is a potential fraud. In summary, without data there would not be a system at all. Therefore, developing the wrappers that extract the data is an extremely important part of the process.

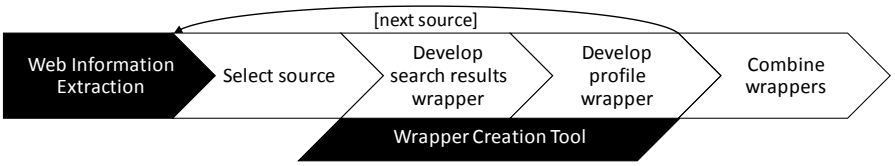

Figure 5.5.: The Web Information Extraction phase of the PSRA Process along with its steps

The entire phase contains four steps, of which three are repeated depending on the number of sources that are included in the system. In the first step, a source is selected from the prioritized list of sources that has been composed in the previous phase. Hereafter, the wrapper for the search results engine of the selected source is developed. The next step is to develop the wrapper for the profiles on the selected source, which will extract the actual data from the specific source. These first three steps are repeated when it is decided that another source will be included in the system. The next source from the list will be selected and, subsequently, both wrappers will be developed. The development of these wrapper can optionally be, partly, automated with the use of Wrapper Creation Tools. After it is decided that no more sources are included at this time, the last step of this phase is executed. In this step the wrappers are combined into one wrapper module, which is also the final deliverable of this phase. The entire phase, along with its steps, is visualized in Figure 5.5.

SELECT SOURCE The first step of the Web Information Extraction phase is a relatively easy step. From the list composed in the previous phase, of which an example can be seen in Table 5.4, the highest ranked source, that is not yet included, is selected. The addition that the source should not have been included yet is because this step, as well as the following two steps, could potentially be executed multiple times. When this step is executed for a second time, the first ranked source, which is Facebook in the example list, has already been included in the system dur-

ing the previous iteration. Therefore, the second ranked source would then be included, which is Twitter in the example list. The wrappers developed in the next two steps will be developed for the source selected in this step.

DEVELOP SEARCH RESULTS WRAPPER The next step in this phase focuses on the development of the search results wrapper. As discussed in Section 3.2.3, public sources such as Facebook contain a lot of data about many different people. Extracting all profiles and their attributes directly from Facebook is an impossible task. As only a very small percentage of the profiles is actually useful for a specific risk analysis, the search engine of the public source is wrapped in order to be able to make a preselection of potentially relevant profiles. As can be seen in Figure 3.10, a wrapper that extracts data from a web source therefore exists of two parts, the search results wrapper and an item wrapper. Since the search results wrapper is the first part of a two step process, it is developed first. Therefore, this step is included tot develop the first part. The second part, the item wrapper, is developed in the next step.

This preselection itself can be done by submitting the personal name of a subject to the search engine of the source. The search results wrapper must therefore be able to accept a personal name as input, and use this as a query for the search engine. The search engine, in turn, returns a list of results based on that particular query. It is the task of the search results wrapper to compose a list of potentially related profiles based on these results. The profile wrapper, developed in the next step, will extract the profiles in this list.

Many public sources, especially the larger ones such as Facebook, offer a service that enables external parties to easily retrieve structured data from their semi-structured web pages. These services are known as Application Programming Interfaces (APIs) and serve as a layer between the internal data storage of the public source and a third party that wants to use the data within that data storage. This way, the owners of the public source can decide which data they want to make publicly available instead of allowing a third party to have full access to their internal data storage. These APIs are easier to wrap than semi-structured web pages as the data is already structured, only the connection has to be initiated.

When a public source does not offer an API, the data has to be extracted via other means. When this is the case, more techniques from the Web Information Extraction research area from Section 3.2 have to be used. For instance, programming languages, specifically designed for extracting data from web pages, can be used to manually develop the wrapper. In addition, extraction rules can also be used to extract the data from the semi-structured web pages. Most of the public sources that contain profiles are generated from templates and filled with data from a database, this means that detail pages for multiple items share the same structural features, and thus the same ex-

traction rules can be used. Therefore, extraction rules have to be defined only once for a particular wrapper.

It was discussed in Section 3.2.2 and visualised in Figure 3.9 that the development of wrappers can also be, partly, automated. These so-called Wrapper Creation Tools utilize machine learning techniques to create a wrapper with various degrees of automation. The degree of automation ranges from supervised to un-supervised, which provide some automation and a lot of automation, respectively. In between, semi-supervised tools exist as well. For an overview of some Wrapper Creation Tools the implementers are referred to Chang et al. (2006).

Although most sources that contain personal data have a search engine to find specific persons on that source, it may occur that a selected source does not have a search engine. When this is the case, their might still be other possibilities to include the source in the system. One option would be to skip the preselection and to nevertheless extract all profiles from that particular source. This might be useful for a smaller source but is not recommended for larger public sources. Another possibility would be to use Google Search as the search engine for a source without its own search engine. Google Search allows to search on a specific site for a given search query, which makes it possible to identify profiles on a site that does not have its own search engine. The *site:* command can be used to limit the search query in Google Search to a specific site. An example search for James Smith on Facebook would look as follows:

James Smith site:http://www.facebook.com/

DEVELOP PROFILE WRAPPER As mentioned above, and depicted in Figure 3.10, a wrapper consists of two parts. After the first part of the wrapper — the search results wrapper — has been developed for a particular source, the profile wrapper for that particular source is up next. The output of a search results wrapper is a list of potentially relevant profiles. The wrapper developed in this step uses that list as input, and returns the structured version of each profile on that list as output. This structured version should at least contain the attributes that were identified in the second phase of the PSRA Process, but other attributes that are available can be included as well.

The development of the profile wrappers is done in the same manner as the search results wrappers in the previous step. The simplest way is to use the API of the particular source to extract structured data of the profile. If such a service is not available, the wrapper can be manually programmed with the aid of programming languages specifically designed for extracting data from web pages. In addition it is also possible to determine extraction rules for the data that should be extracted. Machine learning tools can assist in determining these extraction rules. Using these machine learning tools will result in a certain degree, which depends on the tool, of automation.

COMBINE WRAPPERS After it is decided that no more sources are included at this time, the final step of this phase can be executed. In this step, the wrappers developed in the preceding steps, are combined on two different levels. Firstly, the search results wrappers are combined with their corresponding profile wrapper. This is visually depicted in Figure 3.10. Secondly, all the wrapper sets, each for a specific source, are combined into one wrapper module.

The search result wrapper of each source accepts a personal name as an input and outputs a list of profiles that are potentially relevant to that personal name. Each profile in this list should contain an identifier with which that profile can be uniquely identified on that source, this could be a number, an unique name or the URL of the profile for example. Because the input list contains an unique identifier for each entry, the profile wrapper is able to locate all those entries and extract data thereof. The profile wrapper must accept this list as input and outputs the structured version of each profile on that list. By combining the two wrappers in this manner, a combined wrapper is created that accepts a personal name as input and outputs the structured version of each profile potentially relevant to that personal name.

After all search results wrappers have been combined with their corresponding profile wrapper, these sets of wrappers are combined into one wrapper module. The idea behind the wrapper module is that it accepts a personal name as input, together with a list of public sources where more personal data should be extracted from. The wrapper module should then search for this personal name on all the public sources on the provided list. It does this by utilizing the set of wrappers for each specific source. Each individual wrapper set individually searches the public source it wraps for profiles relevant to that personal name. In the end, the wrapper module combines the structured versions of the profiles from each set of wrappers. This way, the final output of the wrapper module are all the structured versions of profiles extracted from the public sources that were provided, and all relevant to the provided personal name. This wrapper module is the final deliverable of the Web Information Extraction phase.

5.1.5 *Entity Matching*

Entity matching is the fifth, and second to last, phase of the PSRA Process. The Entity Matching research area, discussed in Section 3.3, forms the base for this phase. The aim of this phase is to guide the development of the source-specific matchers. The matchers defined during this phase will decide which extracted profile refers to the same real-world person as the subject of the risk analysis. In order to do this, the attributes extracted by the wrappers, developed in the previous phase, are used. As found during the literature review, similarity functions and decision functions are used to create the matcher. Some of the extracted attributes, such as gender, location and age,

can be used by similarity functions to compare them with the same attributes that are available in the existing internal database. This comparison results in a value that resembles the similarity between the internal attribute and the attribute extracted from a public source. The decision function combines several of these similarity functions in order to make the final decision which profile of that source belongs to the specific subject, or that none of them belongs to the subject. One matcher, the combination of several similarity functions and one decision function, is developed for each source. The combination of these matchers of all sources form the matcher module, which is the final deliverable of this phase.

This phase is part of the PSRA Process for the same reasons as the previous phase, Web Information Extraction. Firstly, the sources differ among the domain and implementation location. Therefore, it might be necessary to develop new matchers for sources that were not included before. Secondly, the matchers developed during this phase are also essential for the overall functioning of the system. Although the wrappers are very good at the job of extracting potentially relevant profiles from the public source, they are not able to determine how relevant those profiles actually are. Let alone that they are able to decide which of those profiles, or none of them, actually belongs to the subject of the risk analysis. When this is not determined and decided upon, the extracted data is useless for the actual risk analysis. After all, it would not make any sense to base a decision related to a person on personal data from another individual. For these two reasons, the development of the matchers are an important phase of the PSRA Process.

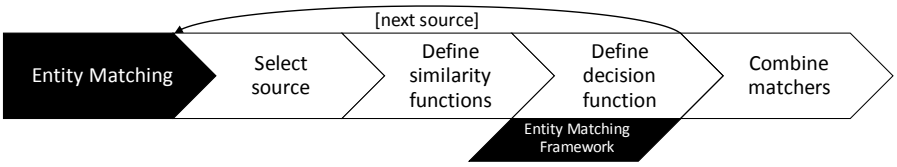

Figure 5.6.: The Entity Matching phase of the PSRA Process along with its steps

The Entity Matching phase, visualized in Figure 5.6, contains four steps in total. Three of these steps are repeated when multiple matchers are being developed. In the first step a source is selected for which the matcher will be developed. The second step aims on the definition of the similarity functions that should be executed on the attributes from the selected source. Hereafter, it is defined how the decision function should combine the similarity functions in order to make a final decision on the extracted profiles. The definition of this decision function can also be partly automated with the use of a Entity Matching Framework. The first three steps are repeated when more sources exist for which wrappers have are already developed, but a matcher has not. When this is the case, the next source will be selected and the similarity functions and decision function will be defined. The final step of this phase is executed when matchers have been defined for all sources wherefore wrappers have been developed. All matchers are combined into a matcher module in this final step.

SELECT SOURCE The Select Source step is a relatively easy step, just as the Select Source step in the previous phase. However, it does differ slightly. Where the Select Source step in the Web Information Extraction phase consisted of selecting the highest ranked source that was not yet included, the step in this phase consists of selecting a source for which wrappers have already been developed, but a matcher has not. Firstly, it makes no sense to define similarity functions and a decision function for a source wherefore wrappers have not been developed. Secondly, it is obviously unnecessary to define these twice. The similarity functions and decision function, which will be defined in the next steps, will be defined for the source selected in this step.

DEFINE SIMILARITY FUNCTIONS The second step in this phase focuses on the definition of the similarity functions that should be executed on the extracted attributes. As found during the literature study in Section 3.3.2, and visually depicted in Figure 3.11, a matcher consists of multiple similarity functions and a decision function. The similarity functions are the first part of this two-step process, and will therefore be defined in this step. The decision function will be defined in the next step.

These similarity functions will calculate a similarity value that resembles the similarity between an attribute extracted from a specific source and that same attribute internally available. So, a similarity function should accept the two values, one from the selected source and one from the existing internal database, of a specific attribute. In general, the more attributes whereon a similarity function is defined, the better. As similarity functions compare the attribute extracted from a specific source, with that same attribute from the internal database, the overlap of attributes between this two sources can be used to define similarity functions on. For each attribute in this overlapping set of attributes, it has to be decided which similarity function is defined. When this has been done for each attribute, the deliverable of this step is completed, namely a set of similarity functions specifically defined for the source selected in the previous step.

As discussed in Section 3.3.2, similarity functions usually return a value between one and zero. Most similarity functions, at least most of the similarity functions encountered during the literature review, are actually string similarity functions such as the Levenshtein (1966) distance. A normalized version of this Levensthein distance, formalized in Algorithm 1, can be used as a similarity function. These can be executed on, for example, the personal name. Whenever a search engine is used to find profiles related to a given personal name, profiles with a different, but similar, personal name are also returned. String similarity functions can be used to calculate the similarity value between the extracted personal name and the internally available personal name. When they are much alike, the string similarity function returns a value close to one, and when they differ much the string similarity functions returns a value closer to

zero. Example similarity functions can be found in Elmagarmid et al. (2007) and Gu et al. (2003).

Other similarity functions can be defined as well. For example, gender can be compared one-on-one. Whenever the gender value extracted from the public source matches the internally available value for gender, a similarity value of 1 is assigned. When they do not match, a similarity value of 0 is assigned. This is formalized in Algorithm 2. Some prepossessing might be necessary as well, as some public sources will return the gender male as a string *male*, whereas the internal source might return the gender male as an integer *1*. The gender extracted from the public source should then first be converted to the format of the internal source to allow a one-on-one comparison. This type will be discussed by example in Chapter 6.

Algorithm 2 Exact match similarity function

```
1: function exactMatchSimilarity(Object object1, Object object2)
2:     if object1 = object2 then
3:         return 1
4:     else
5:         return 0
```

DEFINE DECISION FUNCTION As discussed above, the decision function is the second part of a matcher. Therefore, after the similarity functions have been defined for the specific source, the definition of the decision function is the next step in the creation of the source specific matcher. This decision function should accept the output of multiple similarity functions of the selected source as input, and based on these it should decide whether a profile, and optionally which profile, belongs to the subject of the risk analysis. The final output of a source specific matcher should be one, or no, profile that belongs to the subject of the risk analysis.

First, it has to be decided which approach for the decision function should be taken for this specific source. As discussed in Section 3.3.2 several approaches for the decision functions have been found, namely: numerical, rule-based and workflow-based approaches. Just as the decision of the similarity functions that should be defined, the decision of the approach that should be taken also highly depends on the specific source and available attributes. When the decision for an approach has been taken, it has to be defined how it should be implemented. When it is decided to take, for example, a workflow-based approach, the workflow itself has to be defined. It could be decided to firstly filter out all profiles for which the gender does not match, and then filter out all profiles that contain a relatively low similarity value for the personal name. This is just an example way in which this step can be taken, it can be taken in many other ways as well.

Entity Matching frameworks, as found during the literature review in Section 3.3.3 and also depicted in Figure 3.11, can be

used to automate some parts of the decision function. When, for example, a weighted average approach for the decision function has been chosen, it can assist in determining the value of the weights. But, in order to do this, an adequate data set is necessary. This data set should firstly contain the attribute values of the profiles extracted from the public source. Secondly, it should also contain the values of those attributes for profiles internally available. Finally it should be known which profile from the public source belongs to which profile internally available. This way, the machine learning algorithms of the Entity Matching Frameworks are able to determine how important each attribute is in determining whether a profile belongs to a particular subject or not. For an overview of some Entity Matching Frameworks the implementers are referred to Köpcke and Rahm (2010).

Whether the manual approach is taken, or an Entity Matching framework is put into action, at the end of this step a decision function should have been defined for the selected source. This decision function should be able to decide whether a particular profile belongs to subject, or not. How the decision function decides this exactly, depends on the type of decision function that has been chosen and how it is defined. When — for example — an weighted average approach has been taken, it calculates the weighted average of the output of the various similarity functions for all profiles. Hereafter, it selects the profile with the highest weighted average as the profile of the subject, but only when it exceeds a certain threshold. How the other approaches — the rule-based and workflow-based approaches — decide, is described in Section 3.3.2.

COMBINE MATCHERS After matchers have been developed for all sources for which wrappers have been developed as well, this final step is executed. This step is focused on the combination of the similarity functions and decision function for each source that together form a source specific matcher. In addition, the set of all these source specific matchers are also combined into one matcher module.

All the similarity functions defined for a specific source accept two values, one from the specific source and one from the existing internal database, of a particular attribute as input. These similarity functions calculate a similarity value between these two values, and output this. The combined output of all similarity functions for the specific source should be accepted as input for the decision function defined for that source. The decision function will, in turn, output one, or no, profile for a particular subject. By combining these similarity functions and decision function in this manner, a matcher is created that accepts structured profiles from a specific source and attribute values of the internal database as input, and outputs one, or no profile, which belongs to the subject of the risk analysis.

After a matcher has been created for each source, by combining the similarity functions with the decision function, these match-

ers are combined into one matcher module. This matcher module should accept a list of profiles extracted from each source, and the internal attribute values of the subject for which those profiles were extracted. The matcher module then utilizes all the source specific matchers to identify one, or no, profile for each source per subject. The matcher module then combines all the identified profiles, which belong to the subjects of the risk analysis, in a list and returns them.

5.1.6 *System Construction*

The final phase of the PSRA Process is the System Construction phase. In this phase the components, developed during the previous phases, are combined into the final system. Although the separate components each fulfil an important task, all of them are needed in order for the final system to function correctly. Whether the system functions correctly is also examined during one of the steps in this phase. The final deliverable of this phase, and of the whole PSRA Process as well, is a deployed system that extends a current risk analysis system with personal data from public sources.

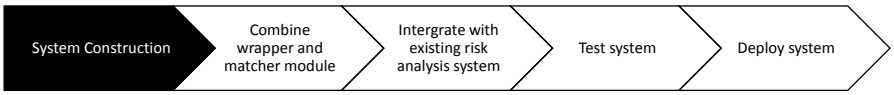

Figure 5.7.: The System Construction phase of the PSRA Process along with its steps

The System Construction phase consists of a fourfold of steps. During the first step the, earlier developed, wrapper module and matcher module are combined to form the final system. Hereafter, that system is integrated with the existing risk analysis system within the organization. In the subsequent step the final system is tested thoroughly. Only when the system successfully passes the testing phase, it is deployed within the organization. This is the last, but foremost step, in which the final result is achieved, a system that extends a current risk analysis with personal data from public sources. The steps of this final phase are depicted in Figure 5.7.

COMBINE WRAPPER MODULE AND MATCHER MODULE In the first step of this final phase the wrapper module and matcher module are combined into the final system.

The wrapper module developed during the Web Information Extraction phase in the PSRA Process accepts a personal name, for example the personal name of the subject stored in the existing internal database, as input. Additionally, a list of supported public sources, from which more personal data should be extracted, is also accepted as input. The wrapper module then searches for personal data potentially relevant to the subject on these sources, and outputs multiple structured profiles for each of these sources. The matcher module, in turn, accepts these profiles per source as input. In addition it also receives the value of the extracted attributes that are also present in the internal database. The matcher module compares each profile

extracted from a public source with the subject profile in the internal data storage. Finally, the matcher outputs one, or no, profile per source that belongs to the subject of the risk analysis based on. In order to decide which profile, or no profile at all belongs to the subject, the decision function within the matcher module is utilized.

By connecting the wrapper module and the matcher module in this manner, the final system is completed. The final system is able to enrich the known personal data of a subject by taking the personal name as input. Additionally, a list of supported public sources, from which the personal data should be extracted, can be provided as input as well. The system then extracts the personal data from these sources and decides what data belongs the the subject of the risk analysis. The output of the final system is one, or no, profile per initially provided source.

INTEGRATE WITH EXISTING RISK ANALYSIS SYSTEM After the final system has been completed, it is integrated with the existing risk analysis in this step. This integration is done in three manners, data from the existing internal database serves as input twice, and data is loaded back into the data warehouse once.

Before the system developed during this process can enrich the personal data about a subject, it requires the personal name of the subject from the existing internal database, such as a Customer Relation Management system, as input. This enables the system to search for relevant profiles on the public sources. A second time the system needs input from the existing internal database, is when the system will determine how relevant a profile is to the subject at-hand. For this, the system needs the internal profile of the subject in order to compare the values of the attributes with those on the public source.

Finally, the output of the developed system is integrated with the existing data warehouse. The output of the system, the extracted personal data from the public sources that belong to the subject of the risk analysis, is loaded into the existing data warehouse. At the end of this step, the architecture of a Business Intelligence system as identified in Section 3.1.2 and depicted in Figure 3.7 and 3.8 is finalized. Thereby, the system has enriched the subject's profile in the existing data warehouse with additional personal data from public sources.

TEST SYSTEM After the system has been completed and integrated with the existing risk analysis system, the system should be tested thoroughly. Although an extensive description on how information systems should be tested in general is outside the scope of this research, a description related to testing some of the specific characteristics will be given below. When implementers do need to acquire general knowledge about testing information system they are referred to the software testing research area.

A valuable way to test the accuracy of the system is to determine the values of the confusion matrix (Kohavi & Provost,

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

Table 5.5.: Confusion matrix (Kohavi & Provost, 1998)

1998). The confusion matrix is depicted in Table 5.5 and contains four values. The value in the upper left box, the *true positive* box, resembles the amount of profiles for which the system correctly decided that a profile belongs to a particular subject. The value in the lower right box, the *true negative* box, resembles the amount of profiles for which the system correctly decided that a profile does not belong to a particular subject. As these two boxes collectively contain the amount of profiles for which the system correctly decided whether it belongs to a particular subject or not, the other two boxes collectively contain the amount of profiles for which the system made an incorrect decision. The *false negative* box resembles the amount of profiles for which the system incorrectly decided that a profile does not belong to a particular subject. These incorrect decisions are the less severe of the two types, as the system only fails to add more personal data to the existing system.

The more severe incorrect decisions are the ones in the *false positive* box, which resembles the amount of profiles for which the system incorrectly decided that the profile belongs a particular subject. This is worse than the previous type since the system now adds unrelated personal data to the existing system. This way, an incorrect decision could be made about a particular subject based on personal data that does not belong to that subject. Therefore, ideally, this amount should be zero. As noted before, decisions should never be solely made based on data automatically acquired data. This is also to prevent these kinds of wrong decision from happening. Not that this method of testing requires an adequate data set with personal data about a subject and the profiles that belong to that particular subject.

DEPLOY SYSTEM Once the system has been tested thoroughly, the final step of both the System Construction phase as well as the entire PSRA Process can be executed. Just as the testing of information systems, the successful deployment of information system is also whole research subject on it own. Therefore, this

is also outside the scope of this research. The final goal of this step, and the entire PSRA Process, is a deployed system that extends a current risk analysis system with personal data from public sources.

5.2 ARCHITECTURE

In the previous section the PSRA Process was introduced as an answer to the first research question. The purpose of this process is to guide the implementation of a system that extends current risk analysis systems with personal data from public sources. A high-level reference architecture for the implementation of this system is provided, as an answer to the second research question, in this section. This high-level reference architecture is based on the artifacts identified during the literature review in Section 3. It provides guidelines for which components should be included in the system, and how these components interact with each other. First, the reference architecture of the system will be presented on a high level. Hereafter, the components within the architecture are described separately in dedicated sub-sections.

Figure 5.8 and Figure 5.9 present the, respectively, functional and technical reference architecture of the system on a high level. The reference architecture is depicted on the same level as the general Business Intelligence system architecture identified in Section 3.1.2 depicted in Figure 3.7. In both architectures the sources are depicted on the left side, and on the right side the analysis functionality which works with the existing data warehouse. In between is the PSRA system, that functions as the extract, transform and load process. The first distinction in the reference architecture can be made between components that are part of the system and components that are not part of the system. Components within the depicted box are part of the system, components outside the depicted box are existing components. These are colored black and white, respectively.

The functional reference architecture contains multiple functionalities which will be implemented by parts in the technical architecture which are discussed later. The customer relation management functionality is responsible for maintaining the profiles of the (prospective) customer of a company, and thus the subjects of the analysis. The wrapping functionality is responsible for the extraction of the data from the public sources. The matching functionality is responsible for matching the candidate profiles with the profiles from the customer relation management functionality. And last, but not least, is the analysis functionality which is responsible for the actual risk analysis of the subjects.

As can been seen in the technical architecture, the system interacts with three types of components outside the system itself, one or multiple sources, an existing internal database, and an existing data warehouse. The sources are the public sources, such as Facebook and Twitter, from which the personal data will be extracted that will be included in an existing risk analysis system. These are the sources that are selected in the Source Selection phase of the PSRA Process presented in the previous section. The existing internal database is

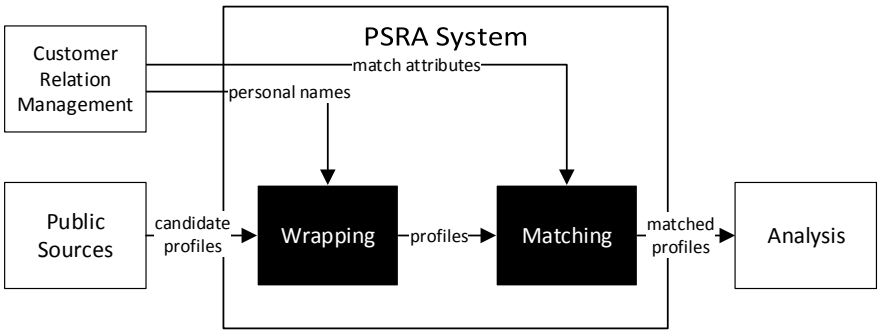

Figure 5.8.: Functional reference architecture of the PSRA System

the database that contains the subjects, and their personal data, of the risk analysis. This is the database of a customer relation management system from the customer relation management functionality in the function reference architecture. Personal data will be loaded from this database. The existing data warehouse is part of the existing risk analysis system of the analysis functionality of the functional reference architecture, data will be loaded into this data warehouse by the system. The integration of this existing internal database, existing data warehouse and the PSRA system itself is done in the System Construction phase of the PSRA Process. The internal components that interact with these external components are shortly introduced below, after which they are extensively discussed in a separate subsection.

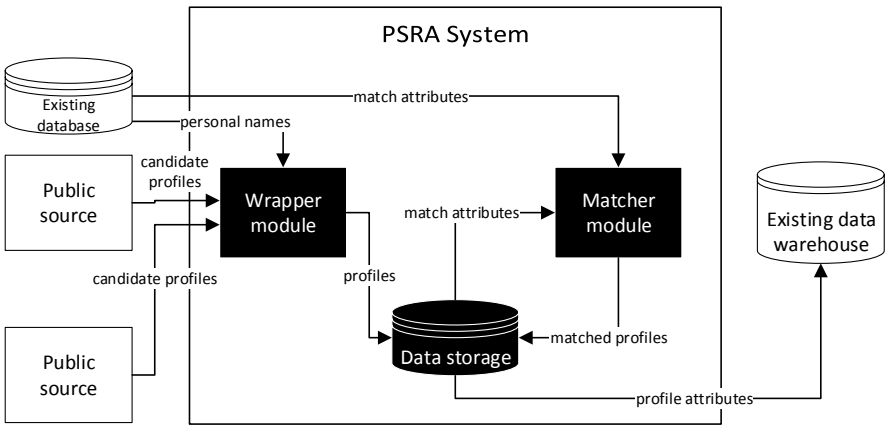

Figure 5.9.: Technical reference architecture of the PSRA System

Three major components can be identified within the system, namely the wrapper module, data storage and matcher module. The wrapper module gathers results from the public sources based on specific requests. These requests are built with the personal names of the subjects of the risk analysis, which are already present in the existing internal database. It should be noted that this could be extended in the future, but for now the search engines of most sources only accept personal names as a query. A future possibility wherein more attributes can be used for the initial search is discussed in Section 6.4. This wrapper module is the technical implementation of the wrapping functionality of the functional reference architecture. The wrapper module's tasks ends when it saves the extracted profiles, which are relevant to the personal names, in the data storage. This data

storage serves as a temporary data storage for the system until it is decided which data belongs to the subjects. This decision is made by the last internal component, the matcher module. The matcher module is the technical implementation of the matching functionality of the functional reference architecture. This module compares attributes from the extracted candidate profiles with the attributes from the existing internal database, and decides which profiles are selected as belonging to a particular subject of the risk analysis. When this is decided, the additional risk attributes, relevant to the subjects of the risk analysis, are loaded from the data storage into the existing data warehouse.

5.2.1 *Wrapper Module*

The first internal component, the wrapper module, is developed during the Web Information Extraction phase of the PSRA Process. This module is responsible for the extraction of the personal data from the public sources. Figure 5.10 and Figure 5.11 zoom in on, respectively, the wrapping functionality and the wrapper module introduced in the high level reference architectures. The functional reference architecture contains two new functionalities, the search results wrapping functionality and the profile wrapping functionality. The search results wrapping functionality is responsible for the extraction of the data from the search results of a public source. The profile wrapping functionality is responsible for the extraction of the data from the profiles of a public source.

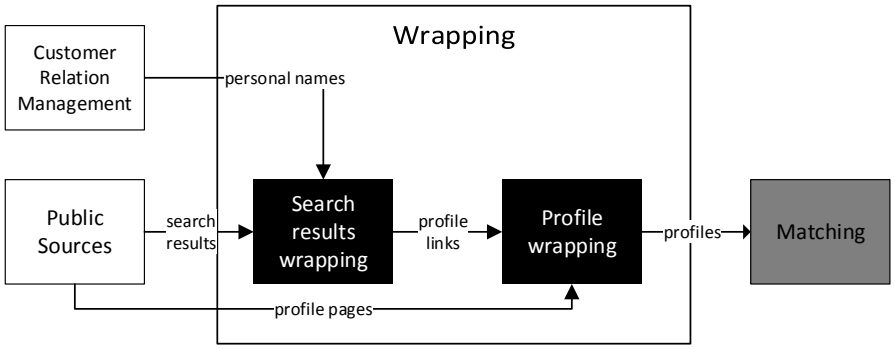

Figure 5.10.: Functional reference architecture of the wrapping functionality

The wrapper module in the technical reference architecture contains two types of components, search result wrappers and profile wrappers. The origin of these two types of wrappers lay within the literature discussed in Section 3.2.3. As extracting all profiles and their attributes directly from a public source is an impossible task, a preselection of potentially relevant profiles has to be made. This is the task of the search results wrapper component. The profile wrapper's task is then to extract these preselected profiles and their attributes from the public source. Only two of these combinations are depicted in the reference architecture, one for each public source included within the system. Of course, more public sources can be included within the system, and thus more than two combinations of these two types of wrappers can exist within the wrapper module.

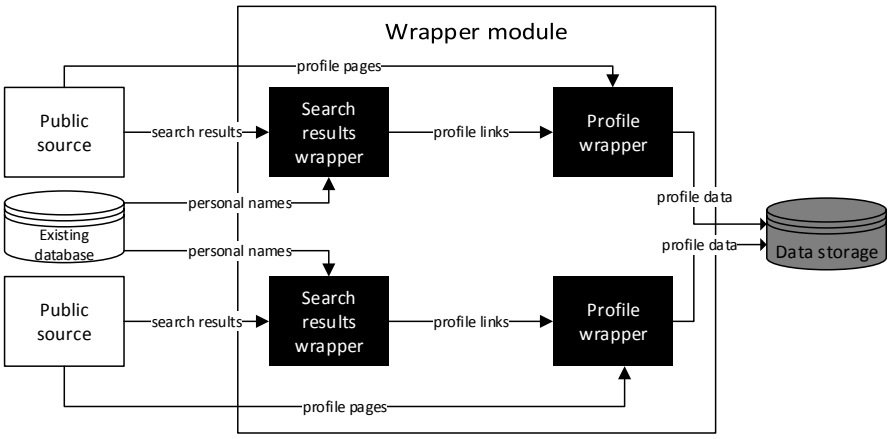

Figure 5.11.: Technical reference architecture of the wrapper module

The search results wrapper is the first link in the chain of extracting the profiles, and their attributes, from the public sources. The search results wrapper is the technical implementation of the search results wrapping functionality of the functional reference architecture. Its task is to make a preselection of potentially relevant profiles. The search results wrapper wraps the search engine of a particular source. It executes search queries based upon the personal names of the subjects present in the existing internal database. The source, in turn, returns the search results of that particular query. The search results wrapper then extracts a list of entries that uniquely identify each profile, for example a link, from those search results.

The task of the profile wrapper is to extract the preselected profile and their attributes from the public source. The profile wrapper is the technical implementation of the profile wrapping functionality of the functional reference architecture. After the search results wrapper has collected the links of the profiles returned by the source's search engine, this list is passed on to the profile wrapper. The profile wrapper component then requests each of these profiles from the source. The source, in turn, returns these profiles to the profile wrapper. The profile wrapper wraps each of these profiles, and extracts the data in these profiles in a structured manner. This profile data is saved into the next component, the data storage.

5.2.2 *Data Storage*

The data storage component is the component that serves as a temporary data storage for the data processed by the system. After the wrapper module has extracted the profiles, relevant to the personal names of the subject, from the public sources, these profiles are stored in the data storage. How the actual data is stored in the data storage, is irrelevant, as long as the matcher module is able to use this data as well. For example, the data stored in the data storage can be stored in a relational database, comma separated files, etcetera. Of course, the decision for a specific type of data storage and the data model used can influence the overall performance of the system. Therefore, careful consideration should go into determining the type and data model.

The next component, the matcher module, will determine for each of the profiles in the data storage whether it belongs to a subject of the risk analysis. If this is the case, the particular profile and their attributes are loaded into the existing data warehouse of the risk analysis system.

Having a separate data storage component, next to the existing data warehouse, can be convenient for several reasons. Firstly, as discussed in Section 3.1.4, a data warehouse contains historical data and once data is entered into the data warehouse it normally is not changed or deleted. The vast majority of the extracted profiles from the public sources probably turn out to be not relevant to subjects of risk analysis. When all these irrelevant profiles are directly loaded into the existing data warehouse, instead of a temporary data storage, the existing data warehouse would contain a lot of unnecessary data that will never be used. In addition, the temporary data storage serves as an interface between the wrapper and matcher module, that is not where the existing data warehouse is intended for.

Secondly, a separate data storage component can prove useful to meet legal obligations. For example, in Section 3.5 one of the obligations extracted from a law was as follows:

> Personal data should not be kept any longer than necessary.

As discussed above, once data is entered into a data warehouse it normally is not changed or deleted. Since this implies that personal data is kept forever, the obligation can not be met. It can be decided to omit this property of the data warehouse, and thus delete the personal data. However, the data warehousing processes are probably not designed for this. A separate temporary data storage makes it much easier to delete the personal data. Once it has been determined by the matcher module which profiles belong to the subjects of the risk analysis, these particular profiles are loaded into the existing data warehouse. Hereafter, the entire temporary data storage can be emptied and all personal data will be deleted.

5.2.3 *Matcher Module*

The matcher module is the last internal component of the reference architecture that will addressed. This module is developed during the Entity Matching phase of the PSRA Process. The purpose of the matcher module is to determine which profiles, extracted by the wrapper module, belong to the subjects of the risk analysis. A more detailed version of the matching functionality's functional reference architecture and the matcher module's technical reference architecture, which were abstractly introduced in the high level reference architectures, are depicted in Figure 5.12 and Figure 5.13, respectively. The Functional reference architecture contains two new functionalities. The similarity calculating functionality is responsible for calculating the similarity between profiles from the wrapping functionality and profiles from the customer relationship management functionality. The decision making functionality is responsible for making the

final decision whether a profile belongs to a subject, or not. This is done with advice from the similarity calculating functionality.

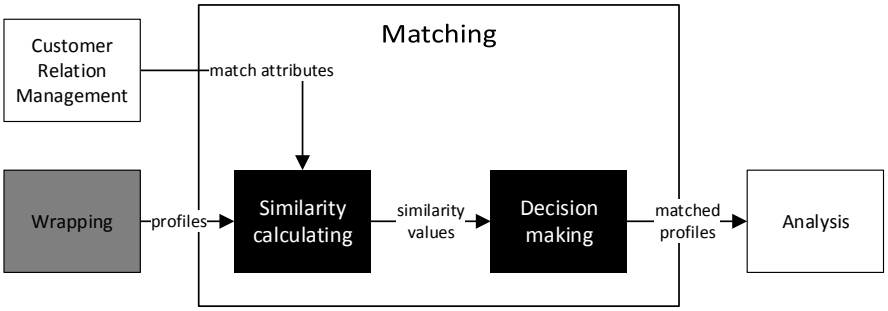

Figure 5.12.: Functional reference architecture of the matching functionality

Just as in the wrapper module, the matcher module in the technical reference architecture contains two types of sub-components. These two types originate from the literature review. To be more specific, from the Entity Matching research area discussed in Section 3.3. The first type of sub components, the similarity functions, are responsible for the calculation of similarity values between profiles. A decision function, the second sub-component type, is responsible for the decision whether a particular profile belongs to a particular subject of the risk analysis. It uses the similarity values computed by the similarity function as a base. Once again, only two of these combinations are depicted in the reference architecture, one for each public source wherefore a wrapper has been developed. When more sources are included within the system, and a wrapper has been developed for these sources, more than two combinations of these two types of sub-components can exist within the matcher module.

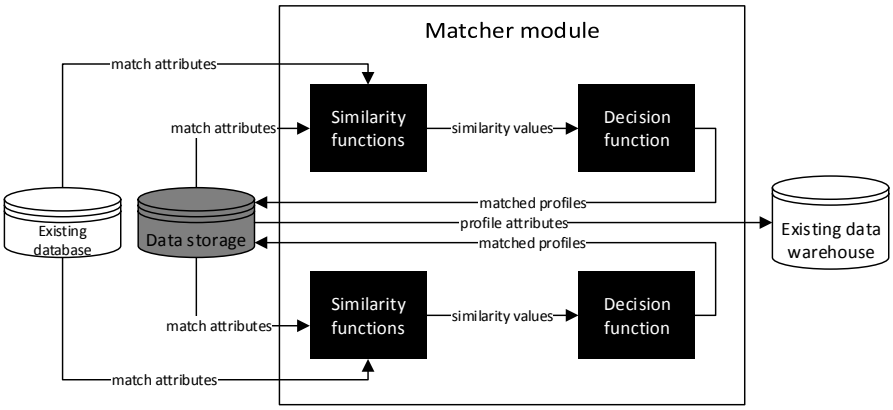

Figure 5.13.: Technical reference architecture of the matcher module

The matcher module starts after the wrapper module has saved the extracted profiles into the data storage. For each origin where profiles have been extracted from, a set of similarity functions have been defined during the Define similarity functions step of the PSRA Process. These similarity functions are responsible for the calculation of the similarity values between profiles and are the technical implementation of the similarity calculating functionality of the functional reference architecture. Based on the origin of an extracted profile, the corresponding similarity functions are executed on this profile.

These similarity functions compare the values of the attributes from the extracted profile, with the values of the same attributes from the subject's profile in the existing internal database. This way, for each attribute of each profile, a similarity value is computed. These similarity values are passed on to the decision function.

The decision function is responsible for the final decision whether a particular profile belongs to a particular subject of the risk analysis. The decision function is the technical implementation of the decision making functionality of the functional reference architecture. The type of decision function, and the specific implementation, have been defined for each origin during the Define decision function step. The similarity values computed by the similarity functions, serve as input for the decision function. The decision function decides, for each specific source and a particular subject, which profile, or no profile at all, belongs that subject. In order to make this decision, it combines the similarity values in a pre-determined manner. As discussed in Section 3.3.2, there are multiple ways how this can be achieved, for example by taken a weighted average. However, regardless of how the decision is made, the profiles that are selected by the decision function are marked in the data storage. In the end, all the marked profiles are loaded into the existing data warehouse.

Part III

EVALUATION

6 EVALUATION THROUGH A PROOF OF CONCEPT

In order to evaluate the PSRA Process and the PSRA Architecture, which were introduced in the preceding part, a proof of concept has been developed. This proof of concept has been developed by, partly, executing the PSRA Process and implementing the system based on the PSRA Architecture. This chapter will firstly discuss how the process was executed, and the problems encountered during that execution. Hereafter, a section will be dedicated to discuss how the PSRA Architecture was used as a reference architecture, and how the system was eventually implemented.

6.1 PROCESS

In Section 5.1 the PSRA Process was presented, which was executed in order to develop the proof of concept. Most of the steps of the process have been taken during this execution, and the experiences gained are presented in this section. Unfortunately, there was no possibility for a case study during the research project. An organization that had the necessary data and systems *and* was willing to participate has not been found. Because of this, some of the steps have not been executed or have been executed slightly different. Although it does provide a first evaluation, multiple case studies should be done in order to fully evaluate the process.

The phases, and their steps, have been executed sequentially and will also be presented sequentially in the following sub section. For each of these phases, and each of their steps, it will be described how they have been executed, which decisions have been made, what the results are and the problems that were encountered during the execution.

6.1.1 *Legal Understanding*

The first phase of the PSRA Process, Legal Understanding, has been executed with the Netherlands as implementation location. This means that the applicable laws and legal obligations were identified for when a system, as proposed in this research, is implemented in the Netherlands. This implementation location has been chosen because of the researcher's good knowledge of the Dutch language, which makes it easier to identify and comprehend the applicable laws.

IDENTIFY APPLICABLE LAWS The applicable laws within the Netherlands have been identified in this step. In order to identify the applicable laws, on-line law books have been utilized. The first applicable law that has been identified is the *The Constitution of the Kingdom of the Netherlands* (2008). This law contains an article, article 10, which is related to the privacy of the citizens. This article has been identified by searching on the website of the European Commission for Democracy through Law. Article 10 also

contains phrases that state that more rules to protect privacy shall be laid down by the Act of Parliament, which implies that — more specific — applicable laws exist. This more specific law has been identified as the second applicable law, and has been found by an ordinary search on the web. This more applicable law is the *Personal Data Protection Act (Unofficial translation)* (2001), which is an implementation of the European directive 95/46/EG (Parliament & the Council of the European Union, 1995). More information on why these laws are applicable can be found in Section 3.5.

The final deliverable of this step in the PSRA process is a list of applicable laws, which was the following list in the case of the proof of concept:

- *The Constitution of the Kingdom of the Netherlands* (2008)
- *Personal Data Protection Act (Unofficial translation)* (2001)

EXTRACT LEGAL OBLIGATIONS The next step that had to be taken was to extract the legal obligations from the laws identified in the previous step. Both identified laws have been examined thoroughly, each article in the applicable parts has been read. For the first identified law, the *The Constitution of the Kingdom of the Netherlands* (2008), this is only one article and no obligations have been extracted from that article. For the *Personal Data Protection Act (Unofficial translation)* (2001), the more specific law, all parts are applicable. Therefore, all articles within this law have been read and each obligation, encountered during the reading, has been marked.

After going trough all the articles, these marked obligations were summarized in a list. The final deliverable of this step of the process, a list of legal obligations, had thus been made. This list was already presented in Section 3.5.3, and contains the following obligations:

- The subject of the risk analysis should have unambiguously given his consent for the processing.
- Along with the consent it is obligated to inform the data subject for which purpose his personal data is processed. It is only allowed to use their data for this purpose they agreed upon.
- Personal data should not be kept any longer than necessary.
- Appropriate technical and organisational measures to secure personal data against loss or against any form of unlawful processing should be implemented.
- Before putting the system in operation the responsible party must notify the processing to the Data Protection Commission.
- The responsible party must not base any decision, which affects the data subject substantially, solely on personal data that has been automatically processed.

- The responsible party must be able to inform the data subject whether or not personal data related to him is being processed and, if so, provide a summary thereof.
- The responsible party must be able to correct, supplement, delete or block the said data in the event that it is factually inaccurate, incomplete or irrelevant to the purpose or purposes of the processing when this is stressed by the subject.
- No personal data may be sent to a country outside the European Union, other than countries that guarantee an adequate level of protection.

DETERMINE PROCESS REQUIREMENTS After the obligations had been extracted from the applicable laws, the obligations that affect the process have been identified. Each of the obligations from the list that had been composed in the previous step, have been examined one by one whether it affects the process. The following sub-list was composed:

- Appropriate technical and organisational measures to secure personal data against loss or against any form of unlawful processing should be implemented.
- Before putting the system in operation the responsible party must notify the processing to the Data Protection Commission.

Additionally, it has been examined for each obligation what is required to be changed in the process in order to meet this obligation. Therefore, a list has been composed with additional requirements for the process. This list, which is the final deliverable of this step, is as follows:

- The PSRA Process should contain a step, before the system is deployed, wherein appropriate organisational measures should be implemented, in order to secure personal data against loss or against any form of unlawful processing.
- The PSRA Process should contain a step, before the system is able to process personal data, wherein the Data Protection Commission is notified of the future processing.

DETERMINE SYSTEM REQUIREMENTS Just as the process requirements had been determined in the previous step, the system requirements have been determined in this step. Again, each of the obligations from the list that had been composed in the Extract legal obligations step, have been examined one by one. Only those that affect the system have been added to the sub-list this time:

- Personal data should not be kept any longer than necessary.
- Appropriate technical and organisational measures to secure personal data against loss or against any form of unlawful processing should be implemented.
- The responsible party must be able to inform the data subject whether or not personal data related to him is being processed and, if so, provide a summary thereof.

- The responsible party must be able to correct, supplement, delete or block the said data in the event that it is factually inaccurate, incomplete or irrelevant to the purpose or purposes of the processing when this is stressed by the subject.
- No personal data may be sent to a country outside the European Union, other than countries that guarantee an adequate level of protection.

After this sub-list had been created, additional requirements for system have been made in order to fulfill the obligations on the list. The final deliverable of this step, the list with additional requirements of the system, is as follows:

- Personal data should be deleted immediately after it is decided that this data is not related to a subject of the risk analysis. Personal data related to a subject of the risk analysis should be deleted immediately after the decision is made for which the data was extracted
- Personal data should only be accessible by authorized persons of the responsible party, and no one else.
- The system should have a function that returns a full list of personal data within the system, given a particular personal name.
- It should be possible to add, edit and delete personal data stored in the data storage at all times.

6.1.2 *Attribute Identification*

The second phase, Attribute Identification, has been executed slightly different since there was no possibility for a case study. Therefore, there was no specific domain for which the proof of concept had to be implemented. Instead, three different domains have been chosen wherefore the steps have been executed for demonstration and evaluation.

SELECT DOMAIN EXPERTS Three different experts in three different domains have been selected for the Attribute Identification phase. All these experts deal with fraud within their specific domains on a daily basis. As these experts are the same experts as the ones selected for the exploratory interviews, an extensive description of the selected experts, and the domains wherin they are active, can be found in Section 4. All of these experts already carried out manual fraud investigations, which made it likely that they posses useful information. The final deliverable of this result, a list of experts, can be seen in Table 2.1. It should be noted that personal names were excluded from this list, since all the experts wanted to remain anonymous. Normally the list would contain the personal names of the experts.

CONDUCT INTERVIEWS After the domain experts had been selected, the next step was to conduct interviews with these experts. In order to ensure the quality of the interview results, an interview protocol has been created. This interview protocol has proven

to be of great value, since it ensured that all topics had been addressed during the interviews.

However, none of the interviewed experts was able to directly indicate the types of personal data they use from public sources during their manual investigations. All of them indicated that they use public sources to acquire background information of a subject, but they could not indicate the type of information that they consider important to determine whether a subject was a potential fraud. They indicated that this differs from case to case. It has been experienced during the interviews that discussing examples of previous cases, wherein personal data from public sources proved useful, yields the best results. However, the experts were only able to give a small number of examples, resulting in a small number of attributes. More, or longer, interviews might result in a larger amount of examples, and thus more attributes as well.

The final deliverable of this step can be found in Appendix C, namely a set of documents with the interview results. It has been chosen to write a summary of each interview, which has been sent to the corresponding expert for validation.

EXTRACT ATTRIBUTES After the summaries had been written for each interview, these summaries have been examined one-by-one and searched through for potential valuable risk attributes. Since none of the experts was able to directly name the valuable risk attributes, they have mostly been extracted from the earlier discussed examples. Additionally, some attributes have also been extracted from examples of attributes used from other sources. From these internal sources, attributes similar to the attributes available on public sources were already extracted. Therefrom, it can be concluded that these attributes can probably be considered important on public sources as well.

The entire list of attributes that has been extracted during this step is presented in Table 6.1, which was also the final deliverable of this step. The identified attributes from all three domains have been merged into one list since the proof of concept has not been developed for a specific domain. This implies that the attributes on this list are potentially valuable risk attributes for a system that would be active in a combination of these three domains. Whether this approach to identifying valuable risk attributes is effective, is not actually validated in this manner. In order to validate this a data set is needed that contains the attributes as they relate to particular subjects and whether these subjects have committed fraud. Unfortunately, such a data set is not available and thus it can not be validated whether the attributes identified in this manner are indeed the most valuable.

PRIORITIZE ATTRIBUTES The list of attributes, that had been identified during the previous step, has been prioritized in this step. During the interviews, it quickly became clear that the experts were most interested in personal data from public sources that allowed them to check whether the data they already have is

Attribute	Reason
Friends	To discover connections with employees within the company
Posts about moving & student or not	To negate a suspicion based on a address change
All personal data present internally (address, date of birth)	To check whether they are similar, differences are suspicious
Posts (about expensive purchases)	Raises suspicion when receiving social welfare payment
Posts (about traveling abroad)	Should be reported when receiving social welfare payment
Personal data	To find contradictions with the data in the RAO system
Posts (about household)	To verify the specified household information
Photos (of household)	To verify the specified household information
Posts (about daily activities)	To verify a 40 hour work statement
Job	To verify job and employer
Age	To see if the salary is reasonable for the subject
Personal data (address, household)	To find contradictions with the specified information

Table 6.1.: Attributes extracted from the interview summaries

correct. These kind of data have therefore been perceived as the more important attributes. However, comparing these data will almost never result in interesting findings since those same attributes are used to match profiles to a particular subject. Therefor, the values of those attributes on the public sources will match the internal values in most cases. This limitation is further elaborated in Section 6.3. Despite this limitation, these kind of attributes are the more important attributes from a technical perspective as well. Since it would be impossible to match these profiles with particular subjects without this data.

Attribute	Importance value
Address/location	4
Date of birth	4
Posts	4
Employer	2
Occupation	2
Relationship	2
Photos	1
Friends	1

Table 6.2.: Prioritized list of attributes

Apart from these attributes that contain personal data also internally present, other attributes have been identified as well. Especially posts —unstructured free texts added by the owner of the profile — were mentioned more than once. Additionally, the current employer, occupation and relationship status were also mentioned multiple times. Photos and a list of friends were

only mentioned once during the interviews. These occurrences, together with the perceived importance, have been used to prioritize the list of attributes. Note that the unstructured free text attributes have been grouped. Rules, as presented in Section 5.1, have been used to guide the prioritization. Although these rules and the occurrence counts have been perceived useful during the prioritization, it proved to be hard to determine the exact importance value. In retrospect, it might have been a nice addition when the list had been prioritized together with the experts. This point for improvement will be discussed in Section 7. Nevertheless, the result, which is also the final deliverable of this step, is presented in Table 6.2.

6.1.3 *Source Selection*

Source Selection is the third phase of the PSRA process and has been executed as it was originally designed. The attributes and their importance values from the previous phase have been used to determine the most valuable public sources that are available in the implementation location, which is the Netherlands in the case of this proof of concept.

IDENTIFY LOCAL SOURCES First of, as much as possible locally available public sources in the Netherlands had to be identified. In order to do so, the public sources mentioned during the interviews have been used as a start. These were: Facebook, LinkedIn, Twitter and Google+. In addition, additional public sources that were used by the implementer were also added. These were: eBay, Marktplaats, Schoolbank and Foursquare. The final deliverable of this step, a list of locally available public source can bee seen in Table 6.3.

Source (Netherlands)
eBay
Facebook
Foursquare
GooglePlus
LinkedIn
Marktplaats
Schoolbank
Twitter

Table 6.3.: List of locally available public sources identified in the Netherlands

CALCULATE ATTRIBUTE FULFILLMENT After the sources had been identified, the attribute fulfillment scores have been calculated. The matrix, presented in Section 5.1.3, has been used to calculate these scores. The matrix that has been created during the execution of this step, is displayed in Table 6.4.

A problem that has been encountered during this selection is that their is no way to differentiate the score in the degree that a certain source can fulfill a particular attribute. For ex-

Source	Address / location (4)	Date of birth (4)	Posts (4)	Employer (2)	Occupation (2)	Relationship (2)	Photos (1)	Friends (1)	Total	Attribute fulfillment
eBay	4	-	-	-	-	-	-	-	4	20%
Facebook	4	4	4	2	2	2	1	1	20	100%
Foursqaure	4	-	-	-	-	-	-	1	4	20%
Google+	4	4	4	2	2	2	1	1	20	100%
LinkedIn	4	4	4	2	2	2	1	1	20	100%
Marktplaats	4	-	-	-	-	-	-	-	4	20%
Schoolbank	4	-	-	-	-	2	1	1	8	40%
Twitter	4	-	4	-	-	-	1	1	10	50%

Table 6.4.: Matrix with attribute fulfillment scores

ample, take the address / location attribute. Both LinkedIn and Google+ receive the same score for this attribute, although Google+ allows users to enter a far more detailed version of the address / location attribute. LinkedIn only saves this attribute as "[large city] Area, [country]", for example: Amsterdam Area, The Netherlands. Google+, on the other hand, allows to save a more detailed address, up to the street name and house number. With the current method to calculate the attribute fulfillment score, this difference is not reflected in the final score. This point for improvement will be addressed in Section 7

PRIORITIZE SOURCES The attribute score that had been calculated in the previous steps has been used as a base for the prioritization in this step. In addition, the perceived popularity of the public sources has been used as well. Because of this, it has been decided that Twitter is considered more valuable than Google+, although Google+ had a much higher attribute fulfillment score. The final — prioritized — list of sources that has been created for the proof of concept is displayed in 6.5. This is also the final deliverable for this step, and the final deliverable for the Source Selection phase as well.

Rank	Source (Netherlands)
1	LinkedIn
2	Facebook
3	Twitter
4	Google+
5	Schoolbank
6	Foursquare
7	Marktplaats
8	eBay

Table 6.5.: Prioritized list of locally available public sources in the Netherlands

It should be noted that, just as the prioritization of the most valuable attributes, the prioritization of the most valuable sources is not really validated in this manner. It can not be determined whether this approach actually results in a list with the most

valuable sources at the top. In order to do so, the same data set as in the Attribute identification phase is needed; a data set that is — unfortunately — currently not available.

6.1.4 *Web Information Extraction*

The Web Information Extraction phase has been experienced as one of the two most complicated and comprehensive phases in the PSRA Process, together with the Entity Matching phase. Technical knowledge is a must, and due to the various types of sources many different ways to develop the wrappers had to be explored. A lot of problems have been encountered during the development of the wrappers and a lot of experience has been gained along the way.

SELECT SOURCE This step has been executed four times, which resulted in four iterations of the two successive steps. Since the implementation is merely a proof of concept, this was considered a sufficient amount to demonstrate the feasibility of a system such as proposed in this research. The sources have been selected in the order in which they had been prioritized, first LinkedIn whereafter Facebook, Twitter and Google+ followed.

DEVELOP SEARCH RESULTS WRAPPER During this step, the four search results wrappers for the four selected sources have been developed. Although the way wherein the final wrappers have been developed differ among the various public sources, the same approach has been taken to determine the final way in which they have been implemented. First, it has been examined for each source whether they provide an API. Whenever a source does not offer an API, Web Information Extraction techniques are used to extract the search results from the search engine's web page. The development of the search result wrapper for each source will be discussed individually, along with the specific problems that have been encountered during the development.

The first source that has been included, LinkedIn, does offer an API to access their data. However, this API only provides access to a very limited version of the search engine. The LinkedIn API requires a system, which makes use of their API, to authenticate as an user. Once the system is authenticated, it only allows searching for people who are connected to the user as whom the system is authenticated. This limitation causes the API to be unusable for to the purpose of a system as proposed in this research. Therefore, the API of LinkedIn is not used in the search result wrapper. Instead, the web page of the search engine of LinkedIn is used. When one searches for a personal name on LinkedIn, the search engine web page sends a request to the server. This server returns the search results in a structured format, called JSON. From this structured data, the links to the profiles are extracted and passed on to the profile wrapper for LinkedIn.

Facebook, the second source that has been included, offers an API to access their data as well. Just like LinkedIn, Facebook requires the system to authenticate as an user. But, contrary

to LinkedIn, Facebook does allow one to search for profiles in their database to whom the authenticated user is not connected. Therefore, this API has been used to extract the search results from the Facebook data set. From the search results, the unique identifier is extracted for each relevant profile and passed on to the profile wrapper for Facebook.

Twitter is the third source that has been included in the system. Twitter offers, just like Facebook and LinkedIn, an API to access their data. This API also provides functionality to search for a specific person based on their personal name. This functionality is used by the Twitter search results wrapper. Other than all other sources, this search functionality also returns all profile fields of each relevant profile. Therefore, it is not always necessary to pass on the relevant profiles to a Twitter profile wrapper. This is only necessary when all tweets need to be extracted, since only the last tweet is returned by the search functionality of the API.

The last source that has been included in the system is Google+. For this search results wrapper there has also been made use of an API provided by Google+. Just as the Twitter and Facebook API, it provides functionality to search for relevant profiles in their database based on a personal name. Unlike Twitter, it does not return all profile fields and thus the unique identifiers of the relevant profiles are passed on to the Google+ profile wrapper.

DEVELOP PROFILE WRAPPER After a search result wrapper had been developed for a particular public source, a profile wrapper has been developed for that source subsequently. Again, the way wherein the final wrappers have been developed differ among the various public sources. But the approach that has been taken is the same for each source. Whenever the source provides access to a structured version of the profile via an API, this API is utilized. Otherwise, a structured version of the profile is extracted from the web pages of those profiles. Each profile wrapper, and the specific problems that have been encountered during the development, will be discussed individually.

For LinkedIn, the source that had been included first, the profile wrapper extracts the personal data from the profile's web page. As discussed above, the LinkedIn API does not allow to retrieve information about profiles to whom the authenticated user is not connected. The link that is passed on by the LinkedIn search results wrapper is essential for the extraction of personal data from the profiles. When one would copy the link of a particular profile where he is connected to, and send it to a friend that is not connected to that person, only a very limited version of the particular profile would be displayed. However, when the other person searches for that particular profile by submitting the personal name to the search query, a special link is returned by the search engine. With this special link, it is possible to view the full version of that particular profile. In short, LinkedIn allows everyone to see the full version of a particular profile, as long as it has been found through a search on the personal

name. This feature of the LinkedIn search engine is used to extract the full profile of each person.

The actual extraction of the personal data from a LinkedIn profile has been implemented with the use of an extraction rule. A LinkedIn profile page contains all data of that profile, in a structured form, somewhere in the source code of that page. Just as their search engine, this data is structured in the JSON format. An extraction rule is used to extract this piece of data from the source code of the page. Hereafter, the available attributes are extracted from this structured version of the profile.

The profile wrapper of the second included source, Facebook, makes use of the API provided by Facebook. The API is able to return a structured version of a profile, based on the unique identifier of that profile that is passed on from the Facebook search results wrapper. However, the API does not return all publicly available information of a particular profile. Instead, the Facebook API only returns a limited set of data from each profile, it even excludes data that is explicitly made publicly available by the owner of the profile. This means that there is a difference between the data visible when visiting the web page of a profile, and the data visible when retrieving that profile through the API. Therefore, an attempt has been made to extract data from the web pages of the profiles. However, due to profound security measures implemented by Facebook, this has not succeeded. Facebook somehow detects that the profile wrapper was not an actual person that was browsing, and confronted the wrapper with a Captcha. Eventually, the account that was used for authentication was removed. Therefore, the profile wrapper has been reverted to use the API again.

A profile wrapper for the third source, Twitter, has not been developed for the proof of concept. As discussed above, the search functionality of the Twitter API already returned all attribute values needed for the matching process. Therefore, it was not necessary to develop a profile wrapper. However, when one would like to extract all tweets for each profile, the API can be used for that. It offers a functionality to retrieve all tweets of a particular profile based on the unique identifier of that profile.

For Google+, the last source that has been included in the proof of concept, the same story applies as for Facebook. The API does not return all data from a particular profile, not even when it is explicitly made publicly available. Again, there is a difference between the available attribute values when visiting the web page of a profile, and the available attribute values when retrieving that profile through the API. However, the Google+ API only excludes the address of the profile. Partly due the profound security measures implemented by Google+, it has been decided to utilize this API for the Google+ profile wrapper.

COMBINE WRAPPERS Once the previous two steps had been executed four times, once per included source, all the search result and profile wrappers, that had been developed, have been combined into one wrapper module. First, the search result wrap-

per and profile wrapper of each source have been combined into a source specific wrapper. In the case of LinkedIn, this is achieved by passing on the list of special links, which enable to retrieve the full version of a profile, from the search results wrapper to the profile wrapper. The search results wrappers of Facebook and Google+ pass on a list of unique identifiers to their corresponding profile wrappers. As discussed before, only a search results wrapper has been developed for Twitter since that also returns all attributes needed for matching.

Hereafter, these wrapper sets have been combined into the final wrapper module of the proof of concept. This wrapper module can be initialized with a sub-set of all the source specific wrappers, the LinkedIn, Facebook, Twitter and Google+ wrappers. The wrapper module is then able to search for relevant profiles on the specified sources, based on the personal name of a subject.

6.1.5 *Entity Matching*

As mentioned in the previous phase, Entity Matching has been experienced as one of the two most complicated and comprehensive phases in the PSRA Process, together with the Web Information Extraction phase. Although it requires slightly less technical knowledge, the decisions that are made during this phase are crucial for the functioning of the final system.

SELECT SOURCE Just as in the Web Information Extraction phase, this step has been executed four times. For each wrapper that had been developed during the Web Information Extraction phase, a matcher has been developed during this phase. These have been developed in the same order as the wrappers.

DEFINE SIMILARITY FUNCTIONS For each included source, a different set of similarity functions has been defined during this step. This is because not all included public sources have the same attributes available that can be utilized for matching purposes. Although this has resulted in a different set for each source, the same approach has been taking for each source. First, the attributes that are available have been identified for the source. Hereafter, these have been compared with the attributes that were available in the internal database. When an attribute was available in both the public source and the internal database, a similarity function has been defined on that attribute.

For LinkedIn, the following attributes are available: personal name, (global) location, birthday, birthmonth and birthyear. Since a case study was not a possibility, there was no real internal database available. Therefore, an internal database has been created for which it was decided to include all attributes that are also available on the public sources. Therefore, at least one similarity function has been created for each attribute. For the personal name, a string similarity function has been defined, more specifically, the Levensthein distance (Levenshtein, 1966). A normalized version of the Levensthein distance, which was

discussed in Section 3.3 and formalized in Algorithm 1, is used. For the (global) location another string similarity function has been defined, the Gotoh distance (Gotoh, 1982). This similarity function has been chosen since this worked well for determining the similarity between, for example, *Utrecht* and *Utrecht Area, The Netherlands*. Other approaches for this attribute could have been taken as well, such as removing the generic part and using the Levensthein distance or by using a geographical service. This will be discussed in the future extensions section of this chapter.

For the birthday, birthmonth and birthyear, four similarity functions have been defined. These four similarity functions have an increasing degree of preciseness. The first similarity function, an exact match function, has been defined on the birthyear alone. Whenever the birtyear on the public source matches the birthday in the internal database, a similarity value of one is returned, otherwise zero is returned. The second and third similarity function have been defined on the birthyear and the birthmonth, and the birthmonth and birthday, respectively. When both of these match, a similarity value of one is returned, otherwise zero is returned. This second and third similarity function require a higher precision, since the profile should match two values at the same time. Finally, the most precise similarity functions has been defined on all three attributes, requiring all attributes to match.

The following attributes are available on the second included source, Facebook: first name, middle name, last name, gender and locale. It has been decided to exclude locale from the proof of concept, since it returned a regional code for the screen language that has been set by the profile's owner. In order to convert this to a country, a look-up service would have been necessary. It has been decided to exclude look-up services in the proof of concept. On first name, middle name and last name a string similarity, based on the Levensthein distance, had been defined.

Additionally, an exact match function has been defined on the gender attribute. If the gender is the same on the public source as well as in the internal database, a value of one is returned, otherwise zero is returned. This exact match function was already formalized in Algorithm 2. However as mentioned earlier, some preprocessing could be required in some cases. Since Facebook, on the one hand, returns the gender as a string with a value of either male or female and the internal database, on the other hand, stores the gender in the ISO/IEC 5218 format, the value of Facebook has to be converted. The ISO/IEC 5218 format stores the gender as follows: 0 equals not known, 1 equals male, 2 equals female and 9 equals not applicable. So, male is converted to 1 and female is converted to 2 before the exact match function determines the similarity. This preprocessing is formalized in Algorithm 3.

Algorithm 3 Facebook gender similarity function

```
1: function FACEBOOKGENDERSIMILARITY(String facebookGender, Integer internalGender)
2:     convertedGender ← 0
3:     if facebookGender = male then
4:         convertedGender ← 1
5:     else if facebookGender = female then
6:         convertedGender ← 2
7:     return exactMatchSimilarity(convertedGender, internalGender)
```

For Twitter, only full name and location were available as attributes. Besides, the location attribute is a free text field, which implies that users, and applications, can store any kind of information in this field they like. Different types of data have been identified in these field including, but not limited to: city, city and country, geo coordinates and free texts that are intended to be funny (Ẅhere I like it!ÿ). On the full name attribute the Levensthein distance similarity function has been defined, and the Gotoh distance similarity function has been defined on the location attribute.

Finally, on Google+, given name, middle name, family name, location and gender are the available attributes. For the three name variants the similarity function based on the Levensthein distance has been defined again, and for the location attribute the similarity function based on the Gotoh distance. Finally, for the gender attribute, an exact match function, with a conversion to the ISO/IEC 5218 format, has been defined.

DEFINE DECISION FUNCTION After a set of similarity functions had been defined for a particular source, the next step was to define the decision function for that source. When the first decision function, obviously for LinkedIn, had to be defined, it was first chosen to implement an average decision function. This decision function took all the similarity functions in the set, and calculated the average of the similarity values that they computed. However, there was a need to be able to value some attributes more than others. For example, the combination of a correct birthmonth and birthyear had to be valued more than only a correct birthyear. Therefore, it has been chosen to implement a weighted average decision function, which allows to weigh certain attributes more than others. During the definition of the subsequent decision functions for the other sources, the weighted average decision function has been chosen as well. Because the various sources have different attributes available, each weighted average decision function has been defined slightly different. An overview of the weights that have been chosen for LinkedIn can be seen in Table 6.6. A complete overview of all weights can be found in Appendix D.

The weights of the different attributes of the decision functions have been determined by trial and error. If an adequate data

Attribute	Weight
Full name	3
City	2
Birthyear	1
Birthyear + month	2
Birthmonth + day	2
Birthyear + month + day	3

Table 6.6.: Weights for the LinkedIn weighted average decision function

set would have been available, it might have been possible to let an Entity Framework, as introduced in the literature review, determine the weights of the individual attributes for each source. The same applies to the threshold that has been defined. This threshold ensures that only a profile with a certain similarity is matched with a subject, not just the profile with the highest similarity value.

COMBINE MATCHERS After the similarity and decision functions had been defined, they were combined into one matcher per source. This has been achieved by passing on the individual similarity values of each profile, to the decision function. The decision functions then calculates the weighted average of all the individual similarity functions per profile. Finally, it selects the profile with the highest similarity. If this profile does not exceed the threshold, no profile is selected from this source. If it does exceed the threshold, it is decided that this profile belongs to that particular subject.

Hereafter, these matchers have been combined into the final matcher module of the proof of concept. This matcher module is able to decide which of the extracted profiles of a particular source, or no profile at all, belongs to a particular subject. It is able to decide this for profiles extracted from LinkedIn, Facebook, Twitter and Google+.

6.1.6 *System Construction*

As mentioned earlier, a case study was — unfortunately — not among the possibilities. Since this last phase, System Construction, heavily depends on an already existing risk analysis system and a corresponding organisation, this phase has only been executed partly. Although the system has been finalized, it has not actually been integrated with a existing data warehouse and it has not actually been deployed in an organisation. Future case studies could evaluate these parts of the phase as well.

CONNECT WRAPPERS AND MATCHERS During the previous two phases, the wrapper module and matcher module have been developed. In order to construct the final system, these two modules have been connected during this step. In order to achieve this, a data store has been introduced as a temporary place to store the attributes extracted by the wrapper module. The matcher

modules selects the attributes of the profiles from this temporary data store. When the matcher module has made a decision which profile to select from each source for each subject, it marks this in the temporary data storage.

INTEGRATE WITH EXISTING DATA WAREHOUSE Because there was no possibility for a case study, and thus there was no existing data warehouse to integrate with, this step has not been executed. The functionality of the proof of concept ceases after it has been marked which profile has been selected from each source for each subject. When there would have been an existing data warehouse to integrate with, the selected profiles could have easily been found in the temporary data storage and their attributes could have been loaded into the existing data warehouse.

TEST SYSTEM The accuracy of the system is tested by using the confusion matrix introduced in Section 5.1. A data set with 25 subjects has been created manually for this purpose. The subjects within this data set were all known to the researcher. Therefore, it was also known whether or not each of these subjects had a profile on the various public sources, and — if they had one — which particular profile it was. This made it possible to check the output of the system. For the sources where a user authentication was required — Facebook — to extract data, this was done from another profile than the researcher's profile that had no connections with the subjects. Thereby eliminating any influences that the connection between the researcher and the data subjects could cause. This data set is relatively small, and only intended as a preliminary evaluation. As will be discussed in Section 10, a large data set is required in order to fully evaluate the system.

The data set contains the personal names for each of the subjects, as well as some additional information. This additional information includes: first name, middle name, last name, gender, birthday and city. The system has been initiated with this data set, after which it searched on the four included sources for profiles that belong to the subjects. Each profile that has been assigned to a subject by the proof of concept, has been checked manually. The performance of the system on each source will be presented separately.

For Facebook, the result are displayed in the confusion matrix in Table 6.7. The top left box, the true positive box, resembles the amount of profiles for which the system correctly decided that a profile belongs to a particular subject. The bottom left box, the false positive box, resembles the amount of profiles for which the system incorrectly decided that a profile does not belong to a particular subject. This profile was incorrectly assigned, since the profile that was selected had a name that only differed one character from the subject's name, and the same gender as the subject's name. Therefore, the selected profile had a similarity value of more than 0.9. The actual subject itself did not have a Facebook profile.

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	10	3 (+7)	P′
	n′	1	1 (+3)	N′
	total	P	N	

Table 6.7.: Confusion matrix for Facebook in the proof of concept

The two right boxes, contain values in parentheses as well. The proof of concept was unable to determine which profile belonged to a particular subject in 10 cases. This happened because multiple profiles were found with the exact same name and gender. As a result, multiple profiles with the same similarity value were found, and was the system unable to determine which profile belonged to the subject. Unfortunately this is insurmountable, since Facebook — as mentioned earlier — only provides the personal name and gender attributes as public attributes. Therefore, it is impossible for the system to differentiate further among the multiple profiles with the same similarity value since they are completely identical from the outside. In 7 of these 10 cases, the subject did have an account on Facebook, and thus they are considered false negatives. This is because the false negative box resembles the amount of profiles for which the system incorrectly decided that a profile does not belong to a particular subject. In 3 of these 10 cases the subject did not have an account on Facebook, these were considered true negatives. This is because the true negatives box resembles the amount of profiles for which the system correctly decided that a profile does not belong to a particular subject. It should be noted that the 3 original false negatives are caused by strange behaviour of the Facebook API, where some profiles did not show up in the search results even though they do exist on the Facebook site.

For LinkedIn, more attributes were available for the matching process, and thus the variation in similarity values was larger. Therefore, there were only 3 subjects wherefore multiple profiles had the same similarity values and it could not be determined which of those belonged to the subject. Additionally, there were 4 cases of a false positive, caused by the way wherein LinkedIn returns the location, namely the region wherein the profile's owner lives. Using a string similarity function on the region from LinkedIn and the city from the internal database

resulted it wrong decisions. A way in which this could be improved is discussed in Section 6.4.

Source	True positive	True negative	False positive	False negative
LinkedIn	14	1	4	3 (+3)
Google+	5	4 (+2)	1	0 (+13)
Twitter	8	5	6	1 (+5)

Table 6.8.: Confusion matrix values for LinkedIn, Google+ and Twitter

Although Google+ also had location as an available attribute for matching, only a very small amount of profiles had a value for this attribute. Because of this, the same situation as with Facebook occurred. In 15 cases, it could not be decided which profile belonged to a particular subject because there were multiple profiles with the same given, middle and family name. In 2 of these 15 cases, the subject did not have a profile on Google+. In the other 13 cases, the subject did have a profile on Google+ and thus these are considered false negatives.

Finally, for Twitter, the location attribute had a value in most of the extracted profiles, therefore the amount of subjects wherefore it could not be decided which profile belonged to it was relatively small (3 cases). But, due to to the various types of data entered into this field (city, city and country, geo coordinates, etcetera) some profiles that actually belonged to a subject had a lower similarity value than another profile. This occurred, for example, on a subject that entered only his country as the location value on his twitter profile. Another profile, which had the same value for the full name attribute, had filled in a city. This city had a higher string similarity with city in the internal database than that the country name of the actual profile had.

DEPLOY SYSTEM Since their was no organisation to deploy the developed system, this step has not been executed. Additionally, the factors involved in the deployment of an information system within an organisation quite substantial, and thus a research on its own.

6.2 ARCHITECTURE

The PSRA Architecture, presented in Section 5.2, has been used as a reference architecture during the development of the proof of concept. This section addresses how this reference architecture has been used and in which way it was implemented. It will not discuss the architecture on a very low-level, however it will present the techniques that have been used for each component of the proof of concept. It should be noted that this is just one way how the PSRA Architecture can be used as a reference, other implementations are possible as well.

Figure 6.1 depicts the architecture of the proof of concept on a high-level. The figure is similar to the reference architecture presented in Figure 5.9, except that the descriptions of the data flows are omitted

and overlays of the used techniques are added. The data flows are omitted for the sake of clarity, and the overlays are added to visualize which techniques were used for each component. Apart from these differences, the overall architecture of the proof of concept is the same as the reference architecture. It turned out that the original reference architecture has been put together well. The different components were all necessary and function properly. Additionally, the data flows between the components were also implemented according to the reference architecture.

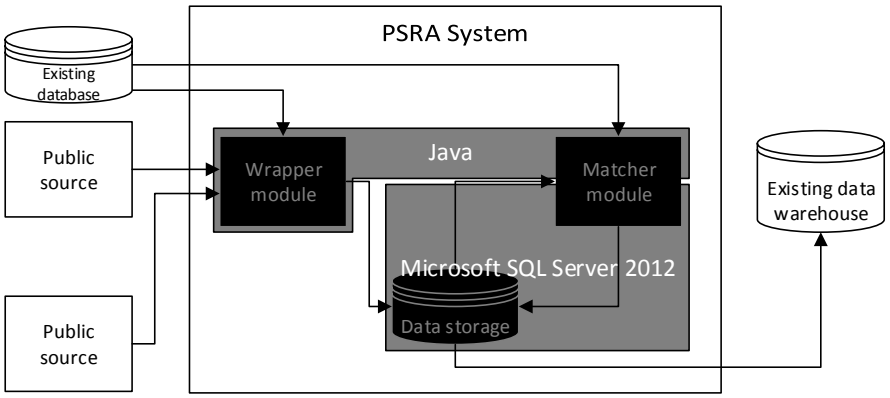

Figure 6.1.: Architecture of the proof of concept

Two different overlays, and thus two different techniques, can be identified in the visualized architecture. For the wrapper module, techniques from Java have been used and for the matcher module techniques from Microsoft SQL Server 2012 have been used. Additionally, both of these techniques have been used for the data storage component. The architecture of the public sources as well as the architecture of the existing data warehouse, depend on the specific source and the data warehouse at-hand. Since there was no existing data warehouse in the proof of concept, the architecture thereof could not be visualized.

Java is a platform independent, object oriented programming language that was used for the development of the wrapper module as well as part of the matcher module. Java was chosen, instead of other programming languages, because of prior knowledge, existing interfaces for APIs of some public sources and the integration possibilities with Microsoft SQL Server 2012.

Micrsoft SQL Server 2012 is a relational database management system that was used as a temporary data storage by the proof of concept. In addition, a part of the matcher module is also developed on the Microsoft SQL Server 2012 database. Microsoft SQL Server 2012 was chosen because of prioir knowledge and the integration possibilities with Java.

6.2.1 *Wrapper module*

As previously described, the programming language Java was used for the development of the entire wrapper module. This implies that java has also been used for the development of each component within the wrapper module, as can be seen by the overlay in Figure

6.2. Just as the visualization of the high-level architecture, this visualization of the lower-level architecture of the wrapper module also omits the names of the data flows for clarity reasons. The reference architecture for the wrapper module, which was based upon the theoretical literature, proved to work well in practice. Therefore, it does not differ from the reference architecture.

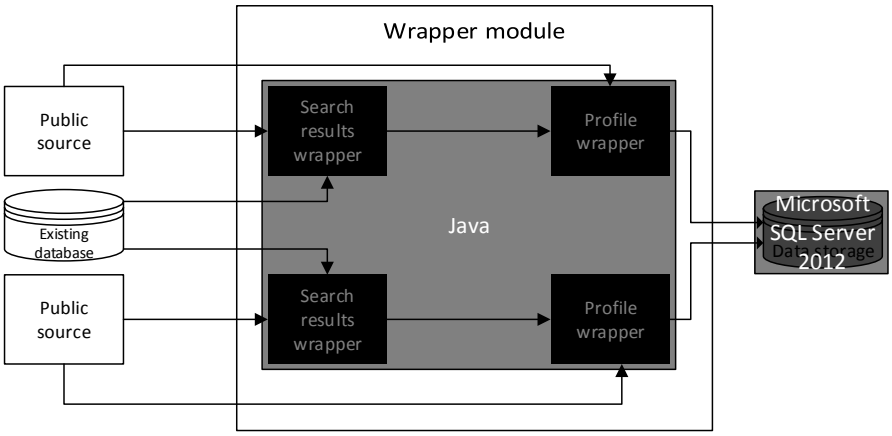

Figure 6.2.: Architecture of the wrapper module of the proof of concept

The distinction between a results wrapper and a profile wrapper worked for almost every source that was included in the proof of concept. The result wrappers in the proof of concept either use the API or, in the case of LinkedIn, the web page of the search engine to retrieve a list of relevant profiles. When the API is used, the result wrapper submits a query to the search function of the API which, in turn, returns a list of relevant profiles. In the case of LinkedIn, a query is submitted to the web page of LinkedIn's search engine. The list of relevant profiles is than extracted from the source code of this page. The only source for which this distinction proved to be unnecessary, was Twitter. The API provided by Twitter already returned all profile information of the relevant profiles. However, when all the tweets are to be extracted from Twitter, a separate profile wrapper would still be necessary to do another request for each profile.

For the actual development of the search result and profile wrappers in Java, third party libraries were used. As indicated above, one of the reasons why Java was chosen was that libraries for APIs of some public sources already existed. These libraries were also developed in Java and may be used freely. They take care of the connection and communication with the public source as well as the authentication on that source. Because the libraries already take care of these aspects, they did not have to be developed during the implementation, which saved valuable time. Instead, work could immediately be started on retrieving the structured data from these libraries. For each source, a distinct search results class as well as a distinct profile wrapper class was developed. The search results class had the task to preselect relevant profiles from the sources, whereas the profile wrapper class collects the structured data from these profiles.

For all sources, except LinkedIn, a third party library for their API was used. However, for LinkedIn a third party client was used to log in onto LinkedIn and extract the data from the web pages with

the use of an extraction rule. As previously discussed, LinkedIn web pages contain all data, specific to that page, in an structured format somewhere in the source code. This structured format is extracted by the third party client based on extraction rules. This structured data can then be used in the rest of the system.

Only a small part of the identified techniques from the Web Information Extraction architecture were needed during the development of the proof of concept. Most sources offered an API through which the major part of the structured data can be retrieved. Extracting the data, which was not available through the API, from their web pages was very difficult due to security measures, which was the case at Facebook and Google+. One technique that was used during the development of the proof of concept were the extraction rules, these were used to extract data from the LinkedIn search and profile pages.

6.2.2 *Data Storage*

It was mentioned in Section 5.2.2 that how the actual data was stored in the data storage, is irrelevant, as long as all the modules are able to access this data. The way in which this is accomplished in the proof of concept is central in this section. Why this way was chosen is also addressed, as well as some recommendations for other possible implementations.

In order to temporary store the data extracted by the wrapper module, a Microsoft SQL Server 2012 database was created. A database was chosen since it allows to easily store and retrieve data, and specifically a relational database because of prior knowledge. Choosing the type of storage was only the first step in the determination of the architecture of the data storage. The way in which the data is stored in the data storage had be to determined as well.

In order to determine the way in which the data would be stored, a data modeling approach was chosen. The data modeling approach that was chosen for the temporary data storage, was the Data Vault data modeling approach coined by Dan Linstedt. This was chosen because some characteristics of the Data Vault modelling approach seemed useful for the development of the proof of concept. These characteristics will be addressed shortly, together with a brief introduction of Data vault modeling. Hereafter, the data model developed for the temporary data storage in the proof of concept will be presented. It should be noted that a complete explanation of Data Vault modelling is out of the scope of this research, interested readers are referred to the excellent book *Modeling the Agile Data Warehouse with Data Vault* by Hultgren (2012).

Data Vault is a data modeling approach used to design the tables for the underlying database of a data warehouse (Hultgren, 2012). A Data Vault data model exist of three types of tables, namely hubs, links and satellites. The hubs are based on core business concepts and only contain business keys. The links are based on a relationship between business keys and only contain this relationship. Finally, the satellites provide a place for all context in a data vault model and contain only that context.

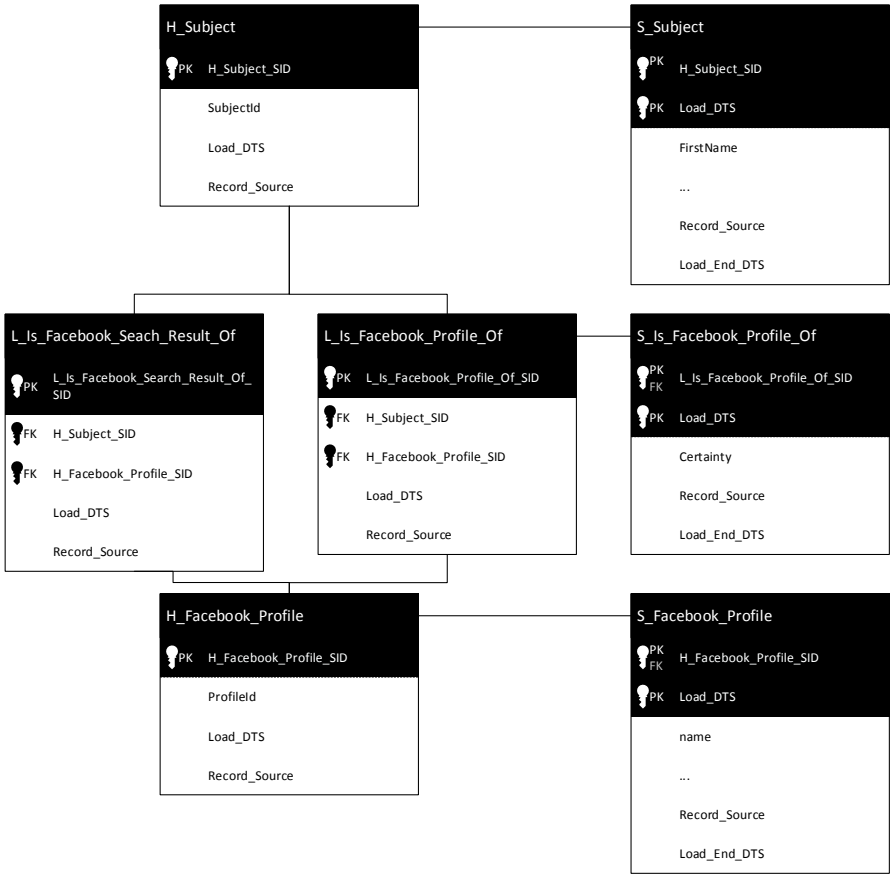

Figure 6.3.: Data Vault data model for personal data from Facebook

Two of the main benefits and specific characteristics of Data Vault data modelling are agility and audibility. These two characteristics were the reason why this approach was chosen. Agility was a convenient characteristic since the proof of concept was developed iteratively. Sources were included in the system one by one, which made it necessary to extend the data model multiple times. The Data Vault data modeling approach supports this since it allows to add new hubs, links and satellites without the need to change the existing data model. The auditability was a convenient characteristic as well since a lot of different components of the system, multiple types of wrappers and matchers per source, loaded data into the data storage. Because of the auditability it was always easy to identify which component loaded the data into the data storage.

Figure 6.3, partly, visualizes the Data Vault data model that was used for the data storage component of the proof of concept. It only contains the hubs, links and satellites related to the storage of the data extracted from Facebook. For the other sources included in the proof of concept, LinkedIn, Google+ and Twitter, similar hubs, links and satellites were modelled as well.

Two hubs can be seen in the data model, namely a Subject hub and a Facebook profile hub. The Subject hub is based on the persons which are the subjects of the risk analyses. The Facebook profile hub is based on the profiles that are extracted from Facebook. Both of these have a corresponding satellite that contains the attributes of each subject and Facebook profile, respectively. The ... indicate that

more attributes are part of the satellite, but those were excluded for the sake of clarity.

Two links are present between these two hubs, one related to the initial search for relevant profiles by the wrapper module and one for the similarity value calculated by the matcher module. First, when the wrapper module has extracted the profiles and their attributes from the public source, these are entered into the database. A link is added between those extracted profiles and the subject where those profiles were a search result of. Second, after the matcher module has determined the similarity between a profile and a subject, a link is added that the Facebook profile is a profile of that subject. Additionally, the certainty of this link is added in the corresponding satellite.

Additionally, in all of the hubs, links and satellites two additional field are modelled: the record source field and the load (end) date time stamp. The record source field is related to the auditability of the Data Vault modelling approach. Whenever an entry is inserted into one of these tables, the auditability field is filled with an unique identifier for the part of the system that inserted the entry. For example, when the Facebook search results wrapper has found a candidate profile for a subject, it inserts an entry into L_Is_Facebook_Search_Result_Of, H_Facebook_Profile and S_Facebook_Profile with the value Facebook Search Results Wrapper in the record source field. The load (end) date time stamp is related to the full historization of the Data Vault modelling approach. Whenever new data is available, for example updated Facebook profile attributes, these are inserted into the data storage with a new load date time stamp, instead of overwriting the previous data. This way, once data is entered into the data storage, it will never be deleted. The load date time stamp field itself enables applications to easily retrieve the most up-to-date values from the data storage, or the values as they were on a specific date.

A relational database proved to work well for the amount of data that the proof of concept had to process. However, when working with extremely large amounts of data, other types of data storage could be a better fit. Implementers that deal with such extremely large amounts of data are recommended to look other options as well. They might be interested in *Big Data* techniques, such as Hadoop and the Hadoop Distributed File System.

6.2.3 *Matcher module*

The matcher module was developed both in Java, the programming language, and in Microsoft SQL Server 2012. The similarity functions that were defined during the PSRA Process were stored in the Microsoft SQL Server 2012 database, and the decision functions who use these similarity functions were developed in Java. This is visualized by two overlays in Figure 6.4. Once again, the data flows are omitted for clarity reasons and the reference architecture proved to work well in practice, therefore there are no differences between the reference architecture and the architecture of the proof of concept.

The similarity functions that were defined during the implementation are stored in the Microsoft SQL Server 2012 database. This

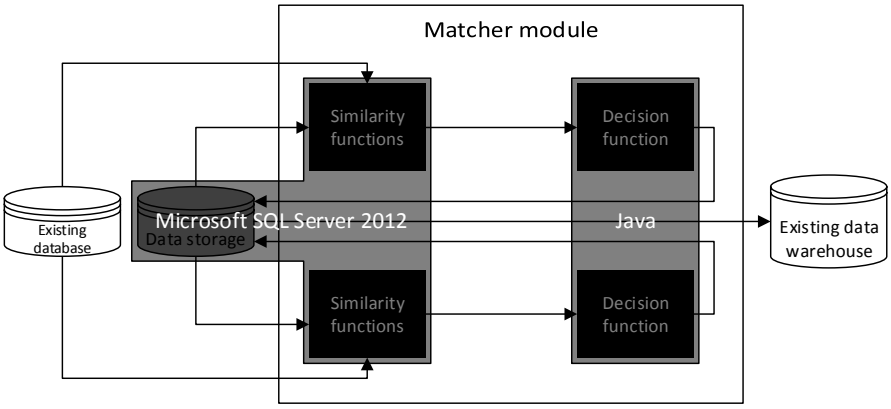

Figure 6.4.: Architecture of the matcher module of the proof of concept

was mainly chosen because of performance reasons, since this way it is avoided that the data should be sent back and forth between the database and the Java program. A third party library is used that contains multiple similarity functions, including the Levensthein distance. Additionally, an exact match similarity function has also been developed for the gender attribute.

Just as the search results and profile wrappers, the decision functions are also developed in Java, which all together form a Java program. In the proof of concept, two types of decision functions have been implemented, an average decision function and a weighted average decision function. For each of these two a distinct class was developed. In addition, one implementation, of one of these two types, was developed for each source. This was done by inheriting the decision function class, and adding source specific information to these classes. For example, the weights of the attributes for the Facebook matcher are set, and thereby an unique decision function class is defined for the source Facebook. The weighted average decision function proved to be the most useful, as it fulfilled the need to be able to reflect that some attributes are more important than others.

Unfortunately, the Entity Matching Frameworks, identified during the literature review, could not be used during the implementation of the proof of concept. As discussed before, Entity Matching Frameworks are able to — partly — automate the definition of the decision function. After it has been decided which approach for the decision function will be taken, the numerical, rule-based or workflow-based approach, the Entity Matching Frameworks can assist in the definition of the decision function. For example, they could determine the weights of each attribute when a numerical approach with a weighted average similarity function has been chosen. However, as also discussed before, this requires an adequate data set. Unfortunately this data set was not available during the implementation, as further discussed in Chapter 9.

6.3 LIMITATIONS

The proof of concept, as discussed in this chapter, has several limitations that affect the potential usefulness of the system. Some of these

limitations can be overcome, others can not. For clarity, a list of limitations that were identified during the development of the proof of concept is presented in this section.

One of the limitations directly related to the proof of concept, as it was implemented during the evaluation, is that the system preselects the profiles solely on the personal name of a subject. As discussed in Section 3.1.3, it is nearly impossible to extract all data from a large public source such as Facebook. Therefore, it is necessary to make a preselection of relevant profile in a manner. However, when only the personal name of a subject, as it is stored in an internal database, is used to make this preselection, a lot of potentially relevant profiles are excluded. Take for example, when a person's name is misspelled, either on the public source or in the internal database. Although some search engines also return similar, slightly different, names, most of these profiles with misspelled names will not be found. In addition, people could also deliberately specify a name different from their actual name, often done from a privacy perspective. They specify a completely different personal name or specify only their initial as first name, which makes it harder to find these people. Section 6.4 presents two extensions to the proof of concept that partially solves this problem.

Another limitation is that not all data of the included sources is available through their API's, not even when the personal data is explicitly made publicly available by the owner of the profile. As an example a profile on Facebook is considered. The owner of this profile has set the privacy setting of his birthday to *Public*. On Facebook, this means that each registered person on Facebook can see his birthday when visiting his profile page, without it begin necessary to be friends with him. However, in order to retrieve his birthday through the API of Facebook, it is necessary to be friends with him. Otherwise, his birthday is not returned by the API. In short, there is a difference between the profile as displayed on the profile page and the profile extracted with the use of the API of Facebook. This is not only the case on Facebook, Google+ has a similar limitation as well, although Google+ only has this limitation on the address attribute. LinkedIn does not allow any information to be extracted with their API, without being friends with the person whose profile is extracted. Of course, this particular limitation can be overcome by extracting the attributes from the profile page, on which all the attributes explicitly made public are viewable. However, this leads to another limitation, related to the policies of the sources included in the proof of concept.

The current implementation of the proof of concept, violates a policy of LinkedIn. This is a major limitation as this will keep the system from being used by companies. The User Agreement of LinkedIn explicitly states that is not allowed to use automated software to extract data from their web pages. So, on the one hand the amount of information that is available is limited when the API is used. On the other hand, extracting the additional data from their web pages is not allowed. This makes it nearly impossible to retrieve the missing public attributes, as other sources have similar policies. It was decided to nevertheless extract data from the web pages of LinkedIn, to demon-

strate that extracting data from web pages is feasible as long as the owner of the source does not disallow it.

6.4 FUTURE EXTENSIONS

Since the system introduced in this chapter is merely a proof of concept, it only serves as a demonstration for the feasibility of the proposed system. Therefore, there is a lot of room for improvement in the future. This section discusses some possible extensions, of which some overcome a limitation discussed in the previous section, on the proof of concept. This is not a complete list of possible extensions, but gives an indication to what could be added in the future.

The first part that could be extended, is the set of similarity functions used by the matcher module of the system. The set of similarity functions in the proof of concept, developed during this research, mostly contains string similarity functions. Although it does contain a wide range of string similarity functions and, in addition, an exact match similarity function for the subject's gender, more similarity functions could be defined.

One similarity function that could be added is a geographical similarity function. On the one hand, some sources return the exact city wherein the subject lives. On the other hand, other sources return a area, such as the state, wherein the subject lives. Currently, in the proof of concept, these are compared using a string similarity matcher, which results in a low similarity value for the city San Francisco and the state of California. However, this similarity value should be relatively high because San Francisco is part of the state of California. An external service, such as Google Maps, could be used to retrieve whether a city is part of a state. It could also determine the distance between two cities, which could reflect the occasions wherein someone lives in a borderland between two cities.

Another similarity function that could be added in the future, is a face similarity function. On most public sources, the picture of a profile is also publicly available. The company that carries out the risk analysis, could also require a photograph of the subject in addition to the personal data they already require. The required photograph and the profile pictures on the source can then be compared in order to determine the similarity.

Another part that could be extended, is the preselection of the relevant profiles. At this moment, the system preselects relevant profiles solely based on the personal name. As discussed in the previous section, this excludes profiles that have misspelled names, anonymized names or abbreviated names. Two extensions will be presented that partially solve this problem.

One way in which this problem can be partially solved, is by deliberately modifying the personal name of the subject extracted from the existing data warehouse. The current proof of concept only queries the search engine of a specific source once for each subject's personal name. An extension could be made that queries the search engine of a specific source several times, each time with a different version of the name. For example, the system could deliberately add common spelling mistakes to the personal name in order to retrieve profiles

with misspelled names as well. It could also abbreviate, or even leave out, parts of the personal name. For example, shortening the first name to only the first letter, or removing a middle name completely. This way, more of the previously excluded profiles will be found.

Although the previous extension enables the system to find more profiles with misspelled or abbreviated names, it does not allow the system to find profiles with anonymized names. Another extension can be developed for this, although this is not possible for each source. Some sources, such as LinkedIn and Facebook, have a more advanced search engine than one that only allows to search on personal names. For example, LinkedIn, allows to search for people based on other criteria such as location, current company, industry, past companies and school without specifying a name. This way, it is possible to find profiles that fit the particular profile of a subject, even when he uses an anonymized name.

Facebook offers similar, but more extensive, kind of functionality with their Graph Search. The Graph Search allows to search for people based on all kinds of criteria by submitting a natural language query. In this natural language query, these criteria can be used to find certain people, without referring to them by name. For example, when it is known that the subject works at Microsoft and is somewhere between 20 and 30 years old the following natural language query could be submitted to Graph Search:

> Men who are older than 20 and younger than 30 and work at Microsoft

Criteria that are included in Graph Search are: gender, age range, relationship, languages, religious views, political views, current employer, past employers, school, likes, birth year, lives in, live near, hometown, locations visited and checked-in. It is evident that Graph Search makes it a lot easier to find people, based on their characteristics, without supplying their name. Unfortunately — at the moment of writing — there is no API available for the Graph Search yet. However, when it does come available, it would probably be a good source to add to the system.

7 IMPROVEMENTS

During the implementation of the proof of concept, which partly evaluated the PSRA Process and the PSRA Architecture, several points for improvement have been identified. Additionally, two experts have given feedback on the PSRA Process and PSRA Architecture as well. From this feedback several points for improvement have been identified as well. This chapter will discuss all these points for improvement and address how the PSRA Process and PSRA Architecture could be improved.

The first point for improvement has been identified during the execution of the Attribute Identification phase of the process, the Attribute prioritization step to be more precise. In this step, the attributes that have been extracted from the expert interview results, have to be assigned an importance value. In the original version of the process this activity was performed by the implementation team, based on the occurrence count and perceived importance of each attribute. However, this has been experienced as a difficult activity to perform. Determining the actual importance value of each attribute has proven to be difficult when one does not have an extensive knowledge of the domain.

This point for improvement can be improved in three manners. First, more interviews could be conducted in order to acquire a more extensive knowledge of the domain. When the implementation team has enough knowledge about the particular domain, it should be easier to prioritize the attributes. However, it could take a long time before this level of knowledge is reached. Besides, the knowledge of the domain will not be used during other parts of the implementation. A second option would be to leave the prioritization of the attributes to the domain experts. This way, a lot of time is saved transferring domain knowledge to people that will normally not reuse this knowledge. This could be done collectively during a session wherein all domain experts are present. This way, the experts, under supervision of the implementation team, can discuss the prioritization until they reach consensus. This would be an extension of the already existing Prioritize attributes step.

A third manner in which the Attribute prioritization step can be improved, was noted by one of the expert. By using data mining technique, it is possible to automatically determine the most valuable attributes within a domain. However, an adequate data set is need to do this. Specifically, a data set with subjects and all of the attributes on the one hand, and whether these subjects committed fraud on the other hand. With this data set, data mining techniques could determine which attributes are the most valuable, and how valuable they are compared to other attributes. This would actually replace the entire Attribute Identification phase.

The second point for improvement has been identified in the Develop search results wrapper step, during the Web Information Ex-

traction phase. It has been experienced that some sources are harder to include successfully than others. Two things can affect the degree of success in including a particular source. First, not all sources provide access to all the publicly available personal data through their API. Of course, Web Information Extraction techniques can then be used to extract this personal data from the actual web pages. But this results in the second thing that affects the success of the inclusion of a particular source, since some source have implemented profound security measures that make it nearly impossible to extract personal data from these web pages. These two problems result in the fact that one source can be included easier than others, or sometimes not even at all.

In the original version of the process, these two problems are not taken into account during the Source Selection phase. This may result in that a source is selected and ranked highly in the Source Selection phase, but that the attributes that caused this high ranking can not be extracted from the source at all. Thus, perhaps it would have been more efficient to include this particular source after other sources, or to not include it at all. Therefore, an improvement for the process would be to include a step in the Source Selection phase wherein the feasibility of the inclusion of a particular source is examined. How feasible the inclusion of a particular source is, can then be taken into account during the Source prioritization step.

The third point for improvement slightly overlaps with the previous point for improvement. It has been experienced during the Calculate attribute fulfillment step of the Source Selection phase that there is a need to vary the score awarded to a particular source based on the extent to which it fulfils that particular attribute. As an example, for the location attribute LinkedIn only returns the general area wherein the owner of the profile lives, whereas other source return the exact city. Evidently, the fulfilment of the location attribute is less on LinkedIn, but not zero. The need to vary the awarded scores originate from this kind of examples. Therefore, an addition to the Calculate attribute fulfillment step of the Source Selection phase would be to allow to award a partial score to a source.

The fourth point for improvement was given by one of the experts, and is related to the maintenance of the system. He questioned whether it was necessary to periodically re-run the process. Of course, periodically execution of the process enables the identification of, for example, new sources or new valuable attributes within the domain. However, it would superfluous to re-run the entire process from beginning to end whenever, for example, a new source is available. Therefore, the original process already contained the events *new source* and *source changed*. But, from this feedback it became apparent that more events could be introduced in order to capture more, preferably all, reasons for maintenance. Examples of events that could be added are: new applicable laws or changed applicable laws on the Legal Understanding phase, new valuable attributes on the Attribute Identification phase, and changes in existing risk analyses system on the System Construction phase. These would extend the current set of events in the PSRA Process.

Part IV

CLOSURE

8

CONCLUSIONS

This research project was set out to investigate how personal data from public sources could be included in risk analysis systems. Therefore, the following main research question was formulated for this research project:

> How can personal data from public sources be utilized for risk analyses of (prospective) customers and thereby support decision-making in fraud sensitive environments?

In order to find an answer to this main research question, multiple sub questions were formulated. An answer to each of these sub questions was sought during the research project and were presented in the preceding chapters. This chapter will first shortly summarize the key findings of, and answers to, each sub question. Hereafter, the answer to the main research question will be addressed.

The first sub question focused on identifying the steps that should be taken to include personal data from public sources in an existing risk analysis system, and was formulated as follows: "Which steps should be taken in order to include personal data from public sources in a risk analysis system¿'. As an answer to this sub question, the PSRA Process was presented. This process contains all the steps necessary to develop a system that extends an existing risk analysis system with personal data from public sources. The main phases of this process are Legal Understanding, Attribute Identification, Source Selection, Web Information Extraction, Entity Matching and System Construction.

In the Legal Understanding phase, the additional process and system requirements are determined, which ensure that the local legal obligations in the implementation location are met. The Attribute Identification phase contains the steps necessary to identify the most valuable attributes in the implementation domain. In the Source Selection phase, the public sources that are available in the implementation location are identified and, based on the most valuable attributes, the sources most valuable for the implementation domain are selected. The Web Information Extraction phase contains the steps necessary to develop the wrapper module, which is able to extract data from the public sources. In the Entity Matching phase, the matcher module is developed, which is able to determine for each subject which profile — or no profile at all — belongs to that particular subject. The System Construction phase contains the final steps necessary to integrate the developed system with the existing risk analysis system.

The aim of the second sub question was to find out what a good underlying architecture could be for a system that extends current risk analysis systems with personal data from public sources. This sub question was formalized as follows: "Which steps should be taken in order to include personal data from public sources in a risk analysis system¿'. The PSRA Architecture was presented in this research as an

answer to this sub question. This architecture contains all the components necessary to be able to extract data from public sources, match them with the subjects of the risk analysis, and add this to the existing risk analysis system. The main components of this architecture are the wrapper module, a temporary data storage and the matcher module. In addition, the architecture describes the data flows between the mutual components, the public sources and the existing risk analysis system.

The wrapper module extracts profiles with personal data from public sources, based on the personal names of the subjects of the risk analysis that reside in the existing internal database. The wrapper module stores these profiles in the temporary data storage component. The matcher module decides, based on the personal data in the temporary data storage and the existing internal database, which profile — or no profile at all — belongs to a particular subject. This decision is stored, together with the certainty of the decision, in the temporary data storage. The personal data of these selected profiles can then be loaded into the data warehouse of the existing risk analysis system.

The third sub question focused on identifying the legal issues that arise when personal data from public sources is used in risk analysis systems, specifically those implemented in the Netherlands. This research question was explicitly formulated by the business, in the following way: "What are the legal issues that arise when personal data from public sources is used in the Netherlands¿'. A literature study into applicable laws resulted in a list of obligations that should be met. The first and foremost obligation that should be met, is that the subject of the risk analysis should have unambiguously given his consent for the processing. Along with this consent, it is obligated to inform the subject for which purpose his personal data is processed. Additionally, the responsible party must notify the processing to the Data Protection Commission before the system is put in operation. Note that this is not a complete list of issues, this list was presented in Section 3.5.3. This complete list was used in the Legal Understanding phase during the process evaluation as well.

And during that evaluation, the answer to the main research question changed quite a lot. Although the — theoretical — answers to the sub questions promised a good result for the main research questions, practice proved otherwise. The proof of concept was unable to extract a large amount of the public personal data due to restrictions of the public sources. Consequential, it was also unable to accurately decide which profile belonged to a particular subject, largely caused by the limited availability of personal data for the matching process.

Therefor, the answer to the main research question is as follows. Personal data from public sources can presumably be included in risk analysis systems with a system implemented by executing the PSRA Process, and basing the architecture of that system on the PSRA Architecture. However, this is purely based on theoretical foundations and was not proven to work in practice mainly due to limited availability of personal data on the selected sources. There are indications that, with more publicly available personal data, this approach can work. Nevertheless, this should be proven by a future case study.

9 DISCUSSION

Although every effort was made to perform a good research, there are always things that could be done even better. Pointing out these things, helps readers to better value the results of the research project. Additionally, it ensures that other researchers with similar research projects are aware of these things as well and take them into account.

First, a limited amount of experts were interviewed for both the exploratory interviews and the attribute identification interviews. Ideally, more experts should be interviewed in order to have a even more solid foundation for the results thereof. Unfortunately, no more experts were found willing to cooperate with this research project. A case study at a cooperative organization might ease this process, since they can appoint several of their experts that can be interviewed. However, such a cooperative organization was not found for this research project.

Second, not all parts of the process and architecture, and thus the systems itself, are fully evaluated in the research project. As mentioned earlier, a cooperative organization for this research project was not found. Since the aim of the system is to extend an *existing* risk analysis system with personal data from public sources, a part of the PSRA process focuses on the integration with that existing risk analysis system. Because a cooperative organization for a case study was not found, this part could not evaluated. The same applies to the selection of the domain experts, as there was not a specific implementation domain.

In addition, the results of the Attribute Identification and Source Selection phase could not be fully evaluated. Although these phases were executed during the proof of concept implementation, it could not be checked if these phases indeed yielded the most value attributes and sources in the particular domains. However, the entire PSRA Process and PSRA Architecture are evaluated by experts in the field, including these parts. The next chapter, Future Research, addresses these points for improvements, and suggests what should be done to fill these gaps.

10

FUTURE RESEARCH

Despite the fact that this research was carried out as comprehensive as possible, there are a number of directions in which further research is necessary. Due to a lack of data sets, and a cooperative organization, a complete evaluation was not among the possibilities during this research. This section describes the directions that could not be addressed, and a call is made for some particular data sets.

First of all, adequate data sets are necessary in order to fully test, evaluate and validate a system as proposed in this research. Because adequate data sets were not available during the research, it was impossible to test, evaluate and validate some parts of the proof of concept. Therefore, a call is made to owners of such data sets to make it available for research purposes. The data sets which are necessary will now be briefly described.

In the Attribute Identification phase, steps are defined that allow the implementation team to identify the most valuable attributes. However, to validate that these steps actually allow the identification of the most valuable attributes, an adequate data set is needed. Specifically, a data set with subjects and all of the attributes on the one hand, and whether these subjects committed fraud on the other hand. With this data set, it could be determined which of all the attributes are important in determining whether a subject committed fraud. This can then be compared to the most valuable attributes identified by the Attribute Identification steps, thereby validating the steps.

Almost the same data sets is necessary to validate the steps of the Source Selection phase. In this phase, steps are defined that allow the implementation team to identify the most valuable sources. Just as with the validation of the steps in the Attribute Identification phase, a data set is needed by which the most valuable sources can be identified. The result of this can then be compared to the most valuable sources identified by the Source Selection steps. For this, the same data set is necessary, but with the addition to each attribute from which source it is extracted. This would make it possible to validate the Source Selection steps.

Although a relatively small data set was constructed in order to preliminary evaluate the proof of concept as a whole, a larger and more complete data set is necessary to fully evaluate the performance of the proof of concept. This data set should contain subjects and their personal data — such as their personal name, birthday, gender and city — and their corresponding profiles on the public sources. This data set would allow to fully evaluate the performance of the proof of concept, instead of an indication of the performance of the proof of concept.

Additionally, a case study should be conducted to evaluate the parts of the PSRA Process that are not fully executed due to the absence of a cooperative organization. This affected the selection of the

domain experts and the System Construction phase. The selection of the domain experts was different because there was no specific domain. In the System Construction phase it was not possible to integrate with an existing data warehouse and to deploy the proof of concept in an organization. A case study would make it possible to evaluate these parts of the process as they were intended.

11

REFERENCES

Artiles, J., Sekine, S., & Gonzalo, J. (2008). Web people search: results of the first evaluation and the plan for the second. In *Proceedings of the 17th international conference on world wide web* (pp. 1071–1072).

Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., Lehmann, P., & Wien, T. (2005). Web data extraction for business intelligence: the lixto approach. *Datenbanksysteme in Business, Technologie und Web*, *11*, 30–47.

Bloom, B. S., Engelhart, M., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. *New York: Longmans: D. McKay Co*.

Blumberg, R., & Atre, S. (2003). The problem with unstructured data. *DM REVIEW*, *13*, 42–49.

Breslin, M. (2004). Data warehousing battle of the giants. *Business Intelligence Journal*, *7*.

Chang, C.-H., Kayed, M., Girgis, R., & Shaalan, K. F. (2006). A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, *18*(10), 1411–1428.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and olap technology. *ACM Sigmod record*, *26*(1), 65–74.

Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, *54*(8), 88–98.

The constitution of the kingdom of the netherlands. (2008). Retrieved September 1, 2013, from `http://www.government.nl/issues/constitution-and-democracy/documents-and-publications/regulations/2012/10/18/the-constitution-of-the-kingdom-of-the-netherlands-2008.html`

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, *19*(1), 1–16.

Embley, D. W., Campbell, D. M., Jiang, Y. S., Liddle, S. W., Lonsdale, D. W., Ng, Y.-K., & Smith, R. D. (1999). Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, *31*(3), 227–251.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, *162*(3), 705–708.

Gregg, D. G., & Walczak, S. (2006). Adaptive web information extraction. *Communications of the ACM*, *49*(5), 78–84.

Grondwet voor het koninkrijk der nederlanden. (2008). Retrieved September 1, 2013, from `http://wetten.overheid.nl/BWBR0001840`

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: Current practice and future directions. *CSIRO Mathematical and Information Sciences Technical Report*, *3*, 83.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, *28*(1), 75–105.

Hultgren, H. (2012). *Modeling the agile data warehouse with data vault*. New Hamilton.

Inmon, W. H. (2005). *Building the data warehouse*. Wiley. com.

Kimball, R. (2006). *The data warehouse toolkit*. Wiley. com.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning*, *30*(2-3), 271–274.

Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, *69*(2), 197–210.

Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S. (2002). A brief survey of web data extraction tools. *ACM Sigmod Record*, *31*(2), 84–93.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady* (Vol. 10, p. 707).

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline*, *9*, 181–212.

Liu, B. (2007). *Web data mining*. Springer.

Luhn, H. P. (1958). A business intelligence system. *IBM Journal of Research and Development*, *2*(4), 314–319.

Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems*, *13*(1), 177–195.

Parliament, T. E., & the Council of the European Union. (1995). Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities*, 31–50.

Pawar, B. S., & Sharda, R. (1997). Obtaining business intelligence on the internet. *Long Range Planning*, *30*(1), 110–121.

Personal data protection act (unofficial translation). (2001). Retrieved September 1, 2013, from `http://www.dutchdpa.nl/Pages/en_wetten_wbp.aspx`

Power, D. J. (2007, March 10). *A brief history of decision support systems.* `ttp://DSSResources.COM/history/dsshistory.html`.

Sarawagi, S. (2002). Automation in information extraction and data integration. In *Vldb 2002, proceedings of 28th international conference on very large data bases, august 20-23, 2002, hong kong, china.*

Soper, D. S. (2005). A framework for automated web business intelligence systems. In *Hicss '05 proceedings of the proceedings of the 38th annual hawaii international conference on system sciences* (pp. 217a–217a).

Srivastava, J., & Cooley, R. (2003). Web business intelligence: Mining the web for actionable knowledge. *INFORMS Journal on Computing*, *15*(2), 191–207.

Turban, E., Sharda, R., Delen, D., & King, D. (2012). *Business intelligence: A managerial approach* (2nd ed.). Pearson Education.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, *40*(9), 96–99.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Wet bescherming persoonsgegevens. (2001). Retrieved September 1, 2013, from `http://wetten.overheid.nl/BWBR0011468`

Part V

APPENDICES

A

PSRA PROCESS

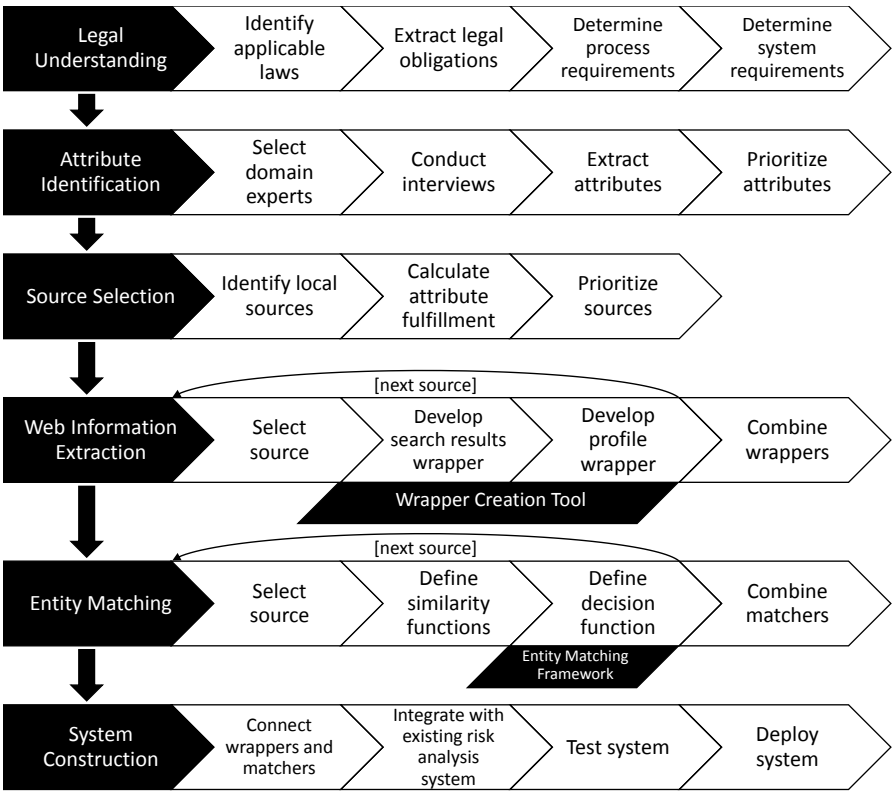

Figure A.1.: The PSRA Process along with all its phases and steps

B

INTERVIEW PROTOCOL

B.1 INTRODUCTION

A moment to mutually introduce each other, the company and the purpose of the interview.

- Could you briefly describe the core business of the company?
- What is your role within the company?
- How long have you been in your present position?
- How long have you been at this company?

B.2 FRAUD

- To what extent does fraud play a role within the company?
- Which parts of the company have to deal with fraud?
- How does fraud affect the overall performance of the company?
- Could you briefly describe your role as it relates to fraud?

B.3 FRAUD DETECTION AND PREVENTION

- Which measures are in place in order to detect or prevent fraud within the company?
- Which manual practices does the company have to detect or prevent fraud?
- Does the company utilize any automatic systems to detect or prevent fraud?
- What kind of automatic system is utilized and how does it work?
- Why is there not an automatic system in place to detect or prevent fraud?
- Would you see any potential in using automatic systems to detect or prevent fraud?

B.4 USE OF PERSONAL DATA FROM PUBLIC SOURCES

- Is personal data from public sources utilized in order to support the detection or prevention of fraud?
- Is this only part of the manual research or also part of the automatic system?

- *Which personal data from public sources do you utilize when manually researching fraud?*
- *Which personal data from public sources is used within an automatic system?*
- *Why specifically these attributes?*
- *Do you feel that some attributes are more important in detecting or preventing fraud than others?*
- Why is personal data from public sources not utilized to support the detection or prevention of fraud?
- Would you see any potential in using personal data from public sources to support the detection or prevention of fraud?

B.5 ETHICAL ISSUES

- Do you have any ethical concerns when utilizing personal data from public sources for the detection or prevention of fraud?
- Does this differ when personal data from public sources is used for other purposes, such as marketing?
- Does it matter to you whether people have explicitly made their data public or that this was the default behaviour?

C

INTERVIEW SUMMARIES

C.0.1 *Interview I*

The company is one of the top 3 e-tailers in the Netherlands and Belgium. Core business of the company is electronic retailing with a focus on consumer electronics. The present function of the expert is manager security and started two months ago at the company. Before starting at this company, he was manager security for two years at another company in the top 3. Focus of the expert is information technology related security, mainly engaged in risk management. His daily task mainly consists of managing security related projects and, during major incidents related to security, he full fills the task of crisis manager. He does not manage a permanent team, but instead forms an team ad hoc for projects of different expertises from departments such as the information technology and finance departments. The expert states that this is because security should be part of the entire company.

According to the expert fraud is present in many parts of the company and committed by both customers and employees. Fraud by employees can take place in the brick store, the warehouse and the different departments such as the information technology and finance departments. Although fraud within the brick store and the warehouse are self-evident, examples indicate that fraud on the information technology and finance departments could also occur. Employees on the finance department could commit fraud, e.g. with gift vouchers and information technology employees could, for example, alter the code of the webshop in their advance.

Fraud by customers can be separated in two groups, normal customers and business customers (B2B). The expert indicates three types of fraud with payment by normal customers, of which two are largely captured by external companies. Using phishing, one could gain access to someones bank account and pay the order from that account, which is mainly the bank's risk. With collect on delivery one could collaborate with the postal worker or steal the delivery, which is mainly the postal company's risk. The third type is related to credit card transactions, with which it is possible to do a chargeback, i.e. charging the money back after the product has been delivered.

The company has various measures in place to detect or prevent fraud, for customers as well as for employees. To prevent employees from committing fraud the principle of least privilege is adopted, which implies that employees can only access the information and resources that they require to fulfill their task. For the first group of customers, the normal customers, multiple measures have been taken. First, they do not allow normal customers to pay on credit, as this payment option is susceptible to fraud. In addition, they also use credit assessments. For business customers they also use the lat-

Attribute	Reason
Friends	To discover connections with employees within the company
Posts about moving & student or not	To negate a suspicion based on a address change
All personal data present internally (address, date of birth)	To check whether they are similar, differences are suspicious

Table C.1.: Attributes identified in Interview I

ter method, acquiring credit assessments to gather information about the financial state of the business customer.

At the moment, no automatic system is in place to support the detection or prevention of fraud. The expert does see potential in a system that assigns risk points to customers, for example when they change their delivery address. These points can be used to indicate possible fraudulent orders.

The company does use personal data from public sources, although this is usually only done when a activity is already flagged as potentially fraudulent. Sites such as Facebook, Twitter, LinkedIn and Google Streetview are mentioned as sources. The expert provides a real-life example from one of the top 3 e-tailers of a customer that redeemed an abnormal amount of gift vouchers, which caused the company to start an investigation. Using the friends list on the customer's profile, they identified a connection between the customer and one of the employees of the company. Further investigation lead to the conclusion that these two collaboratively committed fraud. By this example it becomes evident that the list of fiends of a customer could be a valuable attribute for fraud detection.

Due to the business of the company, e-tailing, the address of the user is also of interest. When a customer changes his address between orders, this can be suspicious. A Facebook post about a movement could explain this sudden change in address. From this example, the expert notes that the information on public sources can be used to *negate* suspicions. In the same context, knowing that the customer is a student could also explain movements. The expert also noted that the personal data that is both present on public sources and internally present is of interest. This can *raise* suspicion when the data specified differs, such as the customer's age.

The expert does not see the usefulness of an automatic system that uses personal data from public sources to support fraud detection yet. He appoints that the data on these sources is diffuse and not well structured. Besides, he thinks that the effort into building such a system can not be justified by the limited use case. He also notes a potential flaw, professional fraudsters could create a fake only identity, thereby circumventing the fraud detection system.

In addition, the expert has some ethical concerns. Using Facebook, for example, for fraud detection is not where it is originally intended for. But on the other hand, he notes that the information is made public by the people themselves and thus it can be used for multiple purposes. He states that it is unethical, and even illegal, to do this for profiling people for marketing purposes. For fraud detection he

thinks it is less unethical, and it even has a positive side for their customers. When a customer tries to pay with a hijacked bank account, fraud detection systems could prevent the withdrawal of money from the original owner of that bank account.

C.0.2 *Interview II*

The expert is analyst at Inspectie SZW and has been working at this company for seven years. Prior to this he worked at the Centrale Justitiële Dienst as a analyst. His daily task consists of supporting projects within the Inspectie SZW by doing analysis based on all sorts of data. In addition to two of these types of projects he is also involved in a criminal investigation and monitoring companies in the chemical sector.

At this moment, the Inspectie SZW has a system in place called Risico Analyse Omgeving (Risk Analysis Environment, RAO). This system extracts data from a great number of systems including, for example, systems of the tax authorities and the municipalities. The system then utilizes this information to carry out risk analyses and presents a list of, for example, potential fraudsters. In order to do this, the RAO system makes use of so-called risk-indicators that are defined by analysts such as the expert. Each indicator has a value that counts towards a score, and persons that have a relatively high score are presented to the inspectors. The persons on this list are then further investigated. The persons that do not stand out based on their score are not accessible by the analysts, thus their identity remains unknown.

According to the expert an example of such an indicator is when someone has a debt at the local tax authorities. All indicators are assigned a value and the height thereof is dependent on how important the indicator is considered. Whenever the indicator applies to a specific person, for example when the person indeed has a debt a the local tax authority, the value of the indicator is added to the total score. A lot of indicators are based on a contradiction between two sources: whenever one source indicates that a person is living alone (making him receive a higher amount of housing benefit) and another indicates he lives in partnership (to receive a high amount of income support) this could indicate that something is not right. All indicators that are invented by the analysts are entered into a database accompanied by a flag whether they have ever been tested or used in a project or if it is only a brainchild. The expert gives an example of a model for a current project, wherein 69 indicators are included. This model has a tested accuracy of 80-85%.

Whenever a project is started, indicators relevant to the project are selected from the database. Additionally, and optionally, the project team organizes a brainstorm meeting with domain experts to invent new indicators related to the project. With this selection of indicators a model specific to the project is built. The model is then used by the RAO system to make a selection of persons that are, for example, potential fraudsters. When a person is selected, the analysts organize a meeting with the involved parties such as, for example, the tax authority and the involved municipality. Each of these in-

Attribute	Reason
Posts (about expensive purchases)	Raises suspicion when receiving social welfare payment
Posts (about traveling abroad)	Should be reported when receiving social welfare payment
Personal data	To find contradictions with the data in the RAO system

Table C.2.: Attributes identified in Interview II

volved parties then check their additional systems, which have not yet been connected to the RAO system, for more signals that increase the suspicion. After this meeting it is decided if the subject is further investigated. According to the expert social media could be considered as an additional party in this meeting, they can then be searched for more signals as well.

Apart from the database with indicators, the analysts also maintain a so-called sourcebook. This sourcebook contains a collection of all internal and external sources that could be used during projects. This sourcebook is invented because it has happened in the past that another team under the same roof had been working with data from a system for a certain period of time. This data was also of value for his team but they did not know about this system. This sourcebook also contains the notion of social media as a source. The expert indicates that extracting the data from some of these sources in the sourecbook takes too long (a period of six weeks is not an exception); when analysts have selected a person based on the data he could have already moved to another location.

The RAO systems as described above currently does not utilize personal data from public sources for the risk analyses. The expert does however indicate that they do use social media after a person has been presented as, for example, a potential fraudster by the RAO system. When doing this, the expert searches for signals that would make the subject even more suspicious. However, social media has a low priority in the projects at Inspectie SZW. Only when the project leader sees clear added value for using social media it is used by the analysts, this does not happen often.

The expert does not see a high potential in using social media as an actual source for the risk analysis of the RAO system itself, as he finds the data on social media too implausible and volatile. People can publish whatever they like on social media and the correctness of these personal data and statements in, for example, posts are not verified. Additionally, a large amount of more credible sources are not yet included in the RAO system, such as the thirteen basic registration systems that will be implemented in the Netherlands. The expert sees more potential in including these sources first after which social media might be added later on. If it is ever added, the indicators that build upon these sources are assigned only a small value to represent the incredibility. The expert also makes the notion of information overload, wherein the analysts receive way too much information to process. Therefore, they do not use all data that is available in the concerned systems but only a selection thereof.

The expert mentions some attributes at which he looks when he investigates a person on social media. First, he searches for contradiction in the personal data in the RAO system and social media. Second, he looks for clues in posts of people. As an example he states that it is at least remarkable when a person who receives social welfare payment posts on his Facebook account that he just bought a new Porsche. However, he immediately indicates that the person in question could also be bluffing to impress friends and thereby notes the incredibility of the information on social media. As another example he mentions a photo album of a vacation to Australia while simultaneously receiving social welfare payment.

Regarding legal issues, the expert indicates that they are required to do a privacy impact analysis for each source that they connect to the RAO system. In addition, they also have to report to the College Bescherming Persoonsgegevens (Dutch Data Protection Authority) at the start of each new project. This authority has also warned them that they did not comply with all the rules. According to the Dutch Data Protection Authority they had not sufficiently described their security plan, they stored personal data too long and they did not sufficiently capture when data was deleted.

The expert does not see ethical problems in using personal data from public sources for their risk analyses. Whenever a user makes use of a service, he should now how to use that service properly. Whenever a person publicly publishes information, he should now that this is public and thus everyone can read and use it, also for the analysts. The expert also does not see ethical problems when the information is published by ignorance, it is the user's responsibility.

C.0.3 *Interview III*

The expert is Fraud Coördinator at a large credit provider in the Netherlands. He has been working for this company for 5 months. Before that he worked as a Fraud Specialist at another credit provider in the Netherlands. He has over 18 years of experience in the field of credit providers. Within the company he is the only employee whose daily tasks are all related to fraud. These daily tasks include, but are not limited to, deciding on possible fraudulent credit requests, investigating fraudulent activities and introducing new measures for fraud prevention.

Fraud plays a major role in the company as each request for credit can potentially be fraudulent. Credit applicants falsify identification documents, payslips and bank statements. According to the expert the falsification of identification documents has significantly reduced due to security features built into the documents. Additionally, credit applicants also deliberately lie about their age, job, household, etc. to influence their financial profile. The expert provides a common example of a credit applicant that has been fired and adjusts the date of his payslip received in the previous month. In addition he also forges the bank statement to reflect the fake payment. Hereby he attempts to receive a credit based on income that he no longer has.

The expert mentions both measures that are introduced at the work place as well as automated measures. At the work place they have an

Attribute	Reason
Posts (about household)	To verify the specified household information
Photos (of household)	To verify the specified household information
Posts (about daily activities)	To verify a 40 hour work statement
Job	To verify job and employer
Age	To see if the salary is reasonable for the subject
Personal data (address, household)	To find contradictions with the specified information

Table C.3.: Attributes identified in Interview III

acceptance team in place that checks every credit application for inconsistencies and decides to accept the credit application or forward it to the expert. They are informed about every aspect of the identification documents, payslips and bank statements such as the security features, font, spacing, lay-out and color of the company, etc. Both the acceptance team as well as the expert also have a *feeling* about certain credit applications due to their experience in the field. According to the expert this *feeling* can not be automated by a system. As an example he describes a credit application that stated that the person was 22 years old and earned €3100 net per month, which caused a feeling that this could not be real.

When a credit application is forwarded to the expert he utilises all kinds of techniques for his research. He contacts banks and employers to verify wage payments. The expert notes that the fraudsters constantly develop themselves and invent new techniques, something that is not possible to detect with automated systems. Additionally, he also communicates with other credit providers to share trends and measures. In this way they try to prevent fraudsters from hopping from one credit provider to the other whenever new measures are introduced.

The company also has an automatic system in place that runs several tests on a credit application in order to detect and prevent fraud. It retrieves the credit applicant's solvency from the central register in the Netherlands, it checks if the identification document is registered in the central register of stolen and lost identification documents, it checks their internal systems for previously registered fraudulent activities and it checks the central system for fraudulent activities in the Netherlands, which is called Externe Verwijzings Applicatie (External Reference Application; EVA). After these tests the company receives the additional documents from the intermediary.

The system also includes a filter functionality wherein the expert can indicate based on prior experience that, for example, all credit applications of people who say that they work at a given company are marked as possibly fraudulent. The tests that are carried out do not only return a positive or negative result, but also search for other indicators of a fraudulent credit application. As an example the expert explains that the central register in the Netherlands also returns the address of a credit applicant, if this differs significantly

from the address of the credit application the system also flags it as possibly fraudulent.

The expert emphasizes that the system never makes a decision on its own. The result, a fraudulent flag or not, is only used as an indication. The reason why the credit application form was flagged as potentially fraudulent is always further investigated. The expert provides an example of the EVA system whereby, when there is a match in the register, he receives the contact information of the company that made the registration. He indicates that he then contacts this company for further information about the registration. Based upon this further research he finally decides whether the credit application is a fraud or not.

During this research the expert indicates that he also uses social media to investigate the subject of the credit application. First, he uses a site that searches the different social media sites for a given name, this reduces the amount of search queries he has to perform. Once the different profiles have been found, he uses the public information on these profiles to get an idea of the type of person. The expert notes that he is always surprised how many information people publicly share. But the expert does agree it is getting less due to attention in the media about privacy matters. As an example he describes an credit application form on which the subject indicated that he is head ICT management at company A. When looking at his profile he sees he is only 24 years old, which seems somewhat young for such a function. But in some cases fraudsters make it even simpler for the expert, they publicly announce their fraud plan on social media including the names of his fellow fraudsters.

According to the expert the extent to which social media is used in the research differs largely among the various subjects. What he looks at on the profiles also differs largely and is highly dependent on the type of fraud, he does not have a uniform way that he follows each investigation. However, he does give two examples of information he could look at during the investigation. First, applicants are required to fill in whether they have a partner and/or children as this is important for the financial profile (less costs). When a subject indicates that he does not have children or a partner but does receive child benefit, he searches for indications of children or a partner in posts and photos. A second example is someone that indicates that he has a full time job but from his posts on social media it becomes clear he spends a lot of time at home or elsewhere. The acceptance team also has Internet access, so when they get the feeling that an credit application could be fraudulent they can also do a parts of this research on social media.

At the moment of the interview the company did not have any systems in place that used social media in an automated manner. According to the expert, especially in his business, this is also not desired. It would slow down the acceptance process because a lot of credit applications would be wrongly flagged as possibly fraudulent. This could happen when the information on social media is not up-to-date and does not match the information on the application form, such as the employer. A lot of fraud investigations would then be started without it being necessary, something the expert could not handle on

its own. The expert mentions multiple disadvantages of using social media in automated systems: the information is often not up-to-date, it lacks a decent history of people as they often switch from social media sites and people could forge their profiles just as they forge their documents. In short, he thinks it is too unreliable.

Just as the output of their system the expert stresses that information acquired through social media should only be used as an indication and further research is always necessary, it should be used as a tool to guide the investigation in a certain direction. The expert does indicate that in an ideal world it would be nice when each credit application would be automatically tested in all manners. However, he indicates that commercial institutions and legislators will very likely prohibit this.

Regarding ethical issues the expert states he does not occupy himself with ethics. He understands that some people have concerns about these kind of systems and he finds it too short sighted to say that people should just pay attention. However, he says that credit providers already have a shortage of options to find out the truth. Therefore, he will not hesitate to use these kinds of ways to investigate fraud as long as it is legal. But again he stresses that it should only be used as an indication, not to solely base a decision on.

D

DECISION FUNCTION WEIGHTS

Attribute	Weight
Full name	3
Location	2
Birthyear	1
Birthyear + month	2
Birthmonth + day	2
Birthyear + month + day	3

Table D.1.: Weights for the LinkedIn weighted average decision function

Attribute	Weight
First name	3
Middle name	1
Last name	2
Gender	2

Table D.2.: Weights for the Facebook weighted average decision function

Attribute	Weight
Full name	3
Location	2

Table D.3.: Weights for the Twitter weighted average decision function

Attribute	Weight
Given name	3
Middle name	1
Family name	2
Gender	2
Birthyear	1
Birthyear + month	2
Birthmonth + day	2
Birthyear + month + day	3
Location	2

Table D.4.: Weights for the Google+ weighted average decision function