Universiteit Leiden

ICT in Business

Method for automated reconciliation of risk rating data
– With a design and implementation at ING

Name: Dimitrios Routsis
Student-no: s1331183

Date: 09/10/2014

1st supervisor: Dr. Emiel Caron (CWTS)
2nd supervisor: Dr. Hans Le Fever (LIACS)

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

First of all, I would like to thank my first supervisor, Dr. Emiel Caron, for his valuable guidance, support and recommendations during my master thesis and my second supervisor, Dr. Hans Le Fever, for assisting me in the initial phase of the master thesis and for his recommendations.

In addition, I would like to thank my manager at ING, Anabel Almagro, for giving me the opportunity to work on this project. Special thanks should be given to two colleagues of mine at ING, Rene Vermaat and Gokulram Krishnamurthy, for their continuous assistance and guidance throughout this project.

Last but not least, I would like to thank my parents, Panagiotis and Ekaterini, and my sister, Maria, for their endless support and encouragement throughout my study.

Business Summary

In this study, a method is designed that automates the integration and reconciliation of risk rating data obtained from credit-rating agencies. Data reconciliation is considered to be part of data integration and, in the business intelligence context, deals with the decision whether data descriptions from different sources point to the same real world entity. Due to syntactic, structural and semantic variation, much of the reconciliation process is done manually, which increases the cost of maintaining data quality.

ING, as many other organizations in the financial sector, is concerned with this problem as it uses risk ratings for risk management. After conducting literature review on data reconciliation and risk ratings, we used ING's resources and risk rating data from risk rating agencies, such as Moody's, Fitch and Standard and Poor's, to identify the main challenges of data integration and afterwards design an automated method for reconciliation. The main focus of this method was the reduction of the manual effort and the securement of the quality of the data. The conceptual design was implemented using ING's infrastructure and evaluated using golden data sets. At the last stage, data quality requirements for onboarding a vendor were identified in order to facilitate the reconciliation process in the future. Our method has very high precision, recall and F1 score values and it matches on average 93.79% issuers.

Keywords: Data Reconciliation, String Matching, Data Cleaning, Object Matching, Data Integration and Financial Data Integration

Table of Contents

1 Introduction

1.1 Background

A big challenge in integrating external data from heterogeneous sources into an internal consistent business database is *data reconciliation*. Data reconciliation, in the context of business intelligence, deals with the decision on whether data descriptions from different sources refer to the same real world entity [Caruso et al. 2000] and is considered to be part of data integration [Lenzerini 2002], which involves the provision of a unique view of the combined data. Due to syntactic, structural and semantic variation, much of the reconciliation process is done manually, which increases the cost of maintaining data quality. ING[1], as many other organizations in the financial sector who buy data from financial institutions, face a similar problem. In order to manage the risk of its business, ING buys risk ratings for various organizations from credit-rating agencies, such as Moody's[2], Standard and Poor's (S&P)[3] and Fitch[4]. The biggest challenge in this process is the development of an automated method to integrate the external with the internal data in order to reduce the manual effort and secure the quality of the data. Consequently, the focus is on developing techniques, methods, algorithms or frameworks for automating the process and successfully integrating the heterogeneous databases. Taking into consideration the new concept of 'big data' [Bizer 2012], the importance of this issue increases.

1.2 Problem description

The main research question of this thesis is:

"How can we design a method that automates the reconciliation of risk rating data obtained from credit-rating agencies?"

In answering the main research question the sub-questions are:

1. What are the main challenges of data reconciliation?
2. How can the method be designed?
3. How can the conceptual model be implemented at ING?
4. How can the method be evaluated?
5. What are the data quality requirements for onboarding a vendor?

[1] http://www.ing.com/en.htm
[2] https://www.moodys.com/
[3] http://www.standardandpoors.com/en_US/web/guest/home
[4] https://www.fitchratings.com/web/en/dynamic/fitch-home.jsp

1.3 Research relevance

1.3.1 Scientific relevance

This research proposal aims at developing a conceptual design and prototypical implementation of a method that takes many aspects of data reconciliation into consideration. In contrast to the majority of the scientific papers, this study will not make any assumptions about the correctness of the data and secondary raw data will be used for developing and validating the method. While various general methods for data reconciliation have been developed, only a small number has focused on external data from financial institutions. The aim of this method is to reduce the manual effort and develop a standardized method for data reconciliation from financial institutions. For the comparison of the statutory legal names of the issuers, two algorithms will be evaluated, Levenshtein's distance algorithm [Levenshtein 1966] and Jaro-Distance algorithm [Jaro 1995], and based on the analysis of the results, the one that provides better ones will be selected. Depending on the business rules, two thresholds will be defined and according to them a matching will be either accepted, or rejected or manually reconciled. The outcome of this method will hopefully have a significant impact on the data quality and integrity. At the last stage, based on the results of the analysis of the data and the challenges of data reconciliation, data quality requirements for onboarding a vendor will be defined.

1.3.2 Organizational Relevance

ING and any other organization in the finance sector can benefit from this method as the manual effort required for the reconciliation process can be decreased. This will not only reduce their operating costs, but it will safeguard the quality and integrity of their data and therefore improve their services. One of the advantages of this method is that it can be easily tuned to the needs and rules of each user as the backbone of the method is standard. Only the inputs and the business rules need to be specified by the user.

1.4 Research methodology

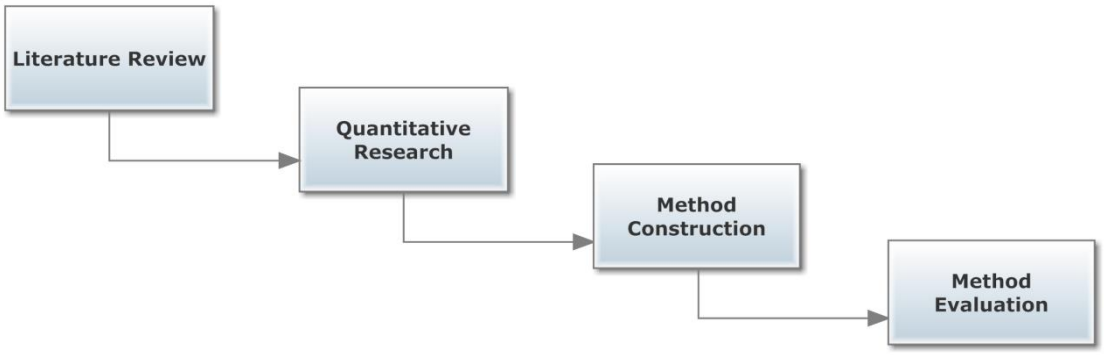
1.4.1 Research design

The research design of our master thesis can be described as causal research. The research problem under scrutiny is very structured and well understood. In addition, for the definition of an automated method for reconciling risk rating data, an analysis of the reasons of mismatches is required. In other words, the causes of the mismatches must be manually defined and afterwards the frequency of each one should be measured. In that way, we will be able to measure the effect of each cause and prioritize them to be resolved.

Consequently, many quantitative methods were used, such as experimentation and simulation. For example, we experimented with many different data cleaning techniques that transform the data in order to match more issuers. On the other hand, the simulation technique was applied at the definition of the threshold for the approximate string matching algorithm. Different sets of thresholds were used and after changing the value of each one, we were calculating the effect on the successful matches and the required manual effort.

1.4.2 Research methods

For the collection of our data, several research method techniques were used. In Picture 1.1, an overview of the research methods that were used is provided.



Picture 1.1 - Research Methods

The literature review process was primarily based on Webster and Watson's (2002) structured approach, which consists of three phases: i) keyword search, ii) backward search and iii) forward search. This approach was preferred than others as it explores in depth the literature background. Furthermore, the literature review was based upon a concept-centric approach as it is more effective and provides higher quality than the author-centric or the chronological-centric [Levy et al. 2006; Webster et al. 2002]. At the first step, the following keywords were selected: Data Reconciliation, String Matching, Data Cleaning, Object Matching, Data Integration and Financial Data Integration. A number of electronic scholarly literature databases were searched on these keywords and the 50 first articles from each, published after 2000, were selected and evaluated. Only the articles that were applicable to the proposed study were included. The backward search step had 3 sub-steps: i) backward references search, ii) backward author search and iii) previously used keywords. The last step, forward search, had 2 sub-steps: i) forward references search and ii) forward author search.

At the second stage, quantitative research, we analyzed secondary data in order to gain a full understanding of the matching challenges and categorize the challenges. The scope of this study is on external sources that provide risk ratings for organizations, such as Moody's, Fitch and S&P. These sources provide daily and monthly files that need to be reconciled. These files contain approximately 10.000 records. The daily files can be characterized as cluster samples, because they are mutually exhaustive and heterogeneous. Therefore, we analyzed a daily file from each of these external sources in order to gain a solid understanding of the matching issues, which assisted us in developing an automated method to resolve them.

Based on the findings, at the Method Construction stage, a conceptual design for data reconciliation was developed. This design was implemented through a standardized data reconciliation interface developed in Microsoft Access using the SQL language (SQL scripts) and in Python scripts, which automatically matched - to certain extent - the external organizations with the internal. A prerequisite for this implementation is data cleansing [Maletic et al. 2010] and transformation [Rahm et al. 2000], and the development and application of some translation rules. At the last stage of this process, Levenshtein's distance algorithm [Levenshtein 1966], implemented in Python scripts, was applied. For the validation of the method, we used a gold dataset (e.g. different Moody's file). For this dataset, a performance measurement technique (e.g. recall, precision) was applied and the accuracy of the method was calculated. After the validation, the percentage of the reduction of the manual effort was estimated, which indicated to what an extent the reconciliation process was automated.

Furthermore, data quality requirements for onboarding a vendor are proposed in order to ensure the quality of data and the reduction of manual effort. For example, if a candidate vendor cannot provide the legal name of the organizations, then ING should not onboard the vendor, because the reconciliation process will not be successful and the internal data quality will be affected.

1.5 Thesis outline

This thesis is organized as follows. Chapter 1 introduces the reader to the context of the research area and describes the problem statement. In order to stress the importance of this topic, the scientific and organizational relevance are defined. Last, the methodology on how this topic was researched is explained by describing the research design and research methods.

Since the problem was how to develop an automated method to reconcile risk rating data, an introduction to the risk ratings was necessary. Therefore, chapter 2 explains what risk ratings are, how are they used and which agencies provide them. Furthermore, we describe how banks use the ratings and more specifically how ING rates its clients by using internal and external ratings.

Chapter 3 describes what data integration is, what general techniques are used to reconcile the external with the internal data and how this process is performed for risk rating data. In addition, the most common challenges of the reconciliation process are described and last, we introduce several approximate string matching algorithms and explain their functionality.

After analyzing the reasons of mismatches, in chapter 4, we designed an automated data reconciliation method of risk rating data. The method consists of the following five steps:

Step 1. Data Cleansing;
Step 2. Match on Cross-reference;
Step 3. Match on Legal Name;
Step 4. Match using BvD Cross-references (only for Moody's);
Step 5. Approximate String Matching Algorithm.

Due to the fact that our method was adjusted by the available infrastructure in ING, we mention some alternative configurations. After designing our method, in chapter 5 we implemented it in order to examine how it works in reality and assess its results. The implementation was done in two main parts. The first one, which was implemented in Microsoft Access, cleans the data and matches the issuers with the internal organization on the cross-reference, legal name and through a third party if it is applicable. The second one, which was implemented in Python, matches the issuers using an approximate string matching algorithm.

Chapter 5 describes the techniques that were used to validate our method. For the validation, a different data set was used than the one in Chapters 4 and 5, in order to avoid the overfitting phenomenon and ensure that our method is valid. In the final part of the thesis, chapter 7, we interpret the results of our research. The main research question and sub-questions are answered and we analyze our main conclusions. Furthermore, the limitations of our research are mentioned and future recommendations are provided.

In Appendix A, the full description of Moody's and Fitch's issuer files is provided. In Appendix B, the table with the results per threshold for Standard and Poor's without the country of residence are provided and in Appendix C, there is the final python script that uses the Levenshtein Distance algorithm including the country of residence to identify and calculate the highest match per issuer.

2 Risk ratings

2.1 Risk ratings

The profitability of commercial and retail banks derives from the difference of the interest on the money that they are lending to other companies, clients or governments and the interest on the money they borrowed. Consequently, in order for banks to be profitable, the interest of the money they lend should be higher than the interest of the money they borrow. This is commonly known as “spread” or “net interest income” [Simpson 2014]. Due to the new regulations and the economic crisis, the net interest income has been reduced and banks are forced to monitor closely their existing clients and evaluate stricter their potential ones [Basel committee 2010]. Thus, they invest heavily in corporate credit risk management (CCRM), which is responsible for monitoring the exposure of the organization to the one-obligor group of customers.

Commercial enterprises and governments in order to raise their capital take loans from banks or develop and sell securities. Securities are financing or investment securities that are sold or bought in financial markets. The most common securities are:

1. Debt securities, such as corporate and government bonds, banknotes and debentures.
2. Equities, such as common stocks.

The main reason why securities are preferred to bank loans is that the borrower does not have to provide huge financial covenants. The company that sells its own securities is known as an *issuer* [U.S. Securities and Exchange Commission 2012]. Typically, an issuer is a corporation, a government or an investment trust and is responsible for the issues. Thus, it should obey all the legal obligations and regulations in jurisdiction [Investopedia 2014]. The credit worthiness of a debtor or an issuer is being evaluated and measured by *credit ratings*. These ratings show the ability of the debtor to pay back its debts and how risky is the dealing with them. One type of credit ratings is bond rating, which measures the risk of default, where the coupon and face value may not be paid back. Bonds are fixed income securities that are used as funds for investments to a firm by investors.

According to Basel II [Basel committee on banking supervision 2014] and III [Basel committee on banking supervision 2010], banks are required to obey to specific regulation in order to be able overcome any potential financial and economic stress, improve their risk management and governance and achieve transparency. One of the three topics that the first pillar of Basel II is focusing is credit risk. Banks should conduct more rigorous credit analysis. As a result, some are using their own credit risk rating system in order to manage properly the externally rated securitization exposure, which is known as the Internal Ratings-Based (IRB) approach [Basel committee on banking supervision 2001]. All the banks that use

this approach should be consistent and show compliance to some minimum requirements. For example, they must provide the risk components that they use and their risk-weight function for these components.

Apart from the internal system, banks can purchase ratings for their potential clients from credit risk rating agencies. There are several risk rating agencies, but the most known ones are Moody's, Fitch and Standard and Poor's (S&P), which are also known as the 'Big Three'. Moody's and S&P are US companies, while Fitch is US-UK. The Big Three agencies control approximately 95% of the rating market share worldwide and they rate the credit worthiness of an issuer. The rating is used for two main reasons. First of all, organizations, such as banks, can evaluate the risk of conducting business with an issuer and therefore decide on whether they will make business with it or not. For example, a bank should only conduct business with issuers that are rated in the three or four highest classes (AAA, AA, A and BBB). The second reason is that this rating directly affects the value of the interest that the issuer should pay. When an issuer is rated as highly risky, i.e. is rated low, then the interest that they security will pay out will be higher [Gitman et al. 2011].

2.2 Risk ratings agencies

There are three big risk rating agencies: Moody's, Fitch and Standard and Poor's (S&P). Moody's and S&P are US companies, while Fitch is US-UK. Even though banks are developing their own credit rating system, they are purchasing credit ratings from external agencies in order to verify the outcome of their system [Treacy et al. 2000]. Some surveys showed that these agencies are adjusting their ratings in a relative slow pace. According to Altman et al. (2004), this is due to the focus of the agencies on the long term and to the fact that they place less weight on short-term indicators to predict the credit quality. Löffler (2005) claims that this occurs because of the informational inefficiency and the rating bounce avoidance. On the other hand, according to Moody's, the focus is on "balancing the market's need for timely updates on issuer risk profiles, with its conflicting expectation for stable ratings" [Cantor 2001].

Another characteristic of the external agencies is the tendency to downgrade countries or issuers more. During the Asian crisis in 1997, the risk rating companies after failing to predict it, they downgraded many countries to an extent that did not reflect their economic situation. The rationale behind that decision was the willingness of the agencies to recover from the damage they made and boost their reputation capital [Ferri et al. 1999]. This decision had big consequences on the economy of these countries, as the cost from borrowing abroad was big. According to C. Kuhner (2001), risk rating agencies have a number of characteristics:

1. They have the power to influence management decisions without being liable and accountable.
2. They avoid assigning different ratings to the same debtor.
3. They can be biased on sociocultural characteristics.
4. If a firm did not request a rating and therefore did not paid for the service, they can rate it assign a lower rate to it.

For the calculation of the rating of an issuer, the risk rating agencies conduct a multidimensional analysis of the company or government. Even though the exact elements of the analysis are unknown and differ between each agency, the investigation covers the legal, financial and management areas. In general, the analysis focuses on the following four characteristics of an issuer [Hawkins et al. 1983]:

- Profitability: This element measures how profitable the firm or the government is and is a strong indicator of the ability of the issuer to meet its obligations. Issuers with stable profitability that have big market share in a stable market in nature are given high ratings. Some common indicators in measuring the profitability are:
    - Net worth to debt;
    - Profit to sales;
    - Earnings variability;
    - Net income to interest;
    - Profit to total assets;
    - Growth rate of earnings per share.
- Liquidity: This element the available cash of the issuer that can be used to pay immediately its liabilities. Some common indicators in measuring the liquidity are:
    - Period of solvency;
    - Working capital to sales;
    - Short term debt to total assets;
    - Cash flow to interest.
- Quality of management: This element measures the ability of the management to meet the goals of the company. Important factors of the management are the corporate strategy, the budget and the vision, because it affects the financial state of the company. Some common indicators in measuring the quality of management are:
    - Years of consecutive dividends;
    - Dividend yield.
- Indenture: This element takes into account the legal aspects of the issue. In the event of a default, indenture provisions may put in a privileged position one party. One common indicator is the subordination status.

2.3 Risk rating issuer files

In the following three paragraphs, we focus on each risk rater and describe the issuer files that they provide to their customers, such as banks. These files can differ from customer to customer depending on the contract. Therefore, the issuer files that are sent to ING are described. These files contain two types of attributes: i) attributes that describe the issuer (e.g. ID, Issuer Legal Name, Country of Residence) and ii) attributes that provide risk ratings.

2.3.1 Fitch's issuer files

Fitch provides daily and monthly issuer files. Every 15 minutes, if there are any changes to the issuers in Fitch's database, an Intra-day file is automatically created and sent to ING through a FTP connection. This file contains only the information for all the issuers whose ratings has changed. On the other hand, if the is no change in Fitch database, no files are sent to ING. The same process is done on a monthly basis. Fitch sends the monthly file with

all the rating information of the issuers that have changed during the previous month. These files are in mir format and contain eighty nine attributes, seven of which are used for reconciliation purpose and eighty one to provide information related to the ratings. From the seven attributes of the file, only for the Issuer ID, Issuer Name, Country Name and Country Code, Fitch is providing information for all the records. Therefore, only these four can be used for the reconciliation process. In Appendix A, more details on Fitch's issuer file are provided.

2.3.2 Moody's issuer files

The primary business of Moody's Investors Service is the analysis of fixed-income securities and debt instruments, and the assignment and publishing of ratings on the creditworthiness of these securities [Moody's Inverstor Service 2009]. Moody's assist investors regarding the ratings, in which they are interested, by publishing them electronically. Furthermore, it offers the Issuer Ratings Delivery Service (RDS – Issuer) to assist in looking at counterparty risk. This service sends via an FTP connection, daily files that contain all the information. The following information is part of the file:

- The Long Term Rating;
- The Rating Outlook;
- The Issuer Rating where available;
- The Estimated Senior Rating;
- The Short Term Issuer Level Rating;
- The Corporate Family Rating (formerly Senior Implied Ratings).

Moody's sends files daily. These files contain all the issuers, whose rating has changed during the previous month or that are new issuers. The file is in txt format and contains forty three attributes. Two of them are used for reconciliation purpose, the unique ID and the legal name, and the rest are used to provide information related to ratings. In Appendix A, more details on Moody's issuer file are provided.

2.3.3 Standard and Poor's issuer files

S&P's system of direct feed is called RatingsXpress. This feed provides daily files in real time and monthly files. Every 5 minutes there is a process that checks if there any changes in Fitch database. If there are, a real-time file is generated in an XML format and is sent to ING. If there are no changes, no file is generated. On the other hand, the monthly master file is compressed using the UNIX standard format (GZIP) and is mainly used for reconciliation. The issuer file contains thirty eight attributes, five of which are used for reconciliation purposes and thirty three for information related to ratings. From the five attributes of the file, only for the Entity ID, Entity Published Name and Country Code, S&P is providing information for all the records. Therefore, only these three can be used for the reconciliation process.

	Fitch	**Moody's**	**S&P**
File Format	.mir	.txt	XML
Delivery Frequency	Daily & Monthly	Daily	Daily & Monthly
Attributes	89	42	38

Table 2.1 - Overview of issuer files

Table 2.1 shows an overview of the issuer files per risk age agency. As it is shown in Table 2.2, the rating scales are very similar. Fitch and S&P have almost the same rating scale. The only difference is that the boundaries of a rating grade can differ a little bit per agency depending on the internal rating system. For example, Fitch's rating for an issuer A can be CCC, while S&P's rating for the same issuer can be CC or B-. Fitch and S&P use a combination of letters and symbols for their ratings, while Moody's uses a combination of letters and numbers.

Moody's	Fitch	S&P
Long-term	*Long-term*	*Long-term*
Aaa	AAA	AAA
Aa1	AA+	AA+
Aa2	AA	AA
Aa3	AA-	AA-
A1	A+	A+
A2	A	A
A3	A-	A-
Baa1	BBB+	BBB+
Baa2	BBB	BBB
Baa3	BBB-	BBB-
Ba1	BB+	BB+
Ba2	BB	BB
Ba3	BB-	BB-
B1	B+	B+
B2	B	B
B3	B-	B-
Caa1		
Caa2	CCC	CCC
Caa3		
Ca	CC , C	CC, C
C	D	D

Table 2.2 - Risk rating scales

2.4 Risk ratings at ING

ING is a global financial institution of Dutch origin, currently offering banking, investment, life insurance (NN Group) and retirement services. ING's Corporate Credit Risk Management (CCRM) team is responsible for providing the platform for the credit approval, credit risk management and reporting of exposures of ING Group, as well as the management of processes supporting these activities like risk research, policy development and systems specification and support. In the CCRM Credit Risk Systems Portal, ING has links to information about the tools and systems it manages. This is an intranet based collection of modules that provide account managers, risk managers and credit analysts with the tools they need to manage and monitor the transaction approval process. Furthermore, it contains useful issue and issuer related financial and market data. The most important characteristic of this portal is the integration of the modules it contains. Access is controlled

by logins tied to viewing rights and user roles, so that it complies with all the ING regulations. The portal consists of: Vortex, GRID, Risk Rater, Financial Statements, Approval Package, Problem Loans, BIR, Loan Pricer, Researcher, Librarian, Market Data, Legal Administrator, CCRM Portal Shared Services and Workflow Services [INGWiki 2014].

ING has its own internal rating system. The account and risk managers can calculate the internal ING risk rating for on organization based on Basel II compliant rating model using the Risk Rater module of the CCRM Credit Risk Systems Portal. The internal ratings are the primary source for evaluating the credit worthiness of an issuer. Nevertheless, ING buys issuer ratings from Moody's, Fitch, and Standard and Poor's in order to supplement its own ratings and reduce the risk of a potential miscalculation.

The Global Relationship Identifier Database (GRID) is a Sybase database that provides a centralized resource for identifying all of ING's customers, and how they are interrelated both legally and economically. Apart from ING's customers, other organizations are stored in order to identify potential customers or monitor the partners of ING's customers to manage the risk. In GRID, many types of data related to organizations are stored such as:

- Legal name of companies;
- Address of companies;
- Customer type;
- Internal risk ratings;
- External risk ratings etc.

It also contains data on the relationship of an organization to other organizations in GRID, namely whether it is a subsidiary, a branch, a fund or legally/economically dependent, i.e. it stores the *Legal Hierarchy* and the *Economic Group Hierarchy*. [Raats 2013] Every record has its own unique GRID ID (8-digits) in order to assist other systems to access them. For example, the company "Example A" has the number "33684582" as a unique GRID identifier (Table 2.3).

GRID ID	Legal Name	Country of Residence	Address of Residence	Town of Residence	…
33684582	Example A	Netherlands	Middelstegracht 8A	Leiden	…
…	…	…	…	…	…

Table 2.3 - Example of a record in GRID

In GRID, there are approximately one million contacts, one million people and eight million legal entities. As a global database, it shares information with hundreds of interfaces around the world, including the domestic banks in the Netherlands, Belgium, and Poland. GRID also informs an array of integrated processes in Vantage, such as those for Customer Support, Relationship Management, CDD, FATCA, and more. This means that all ING entities and business units can rely on each other's results: there is no need for two business units to perform a CDD on the same party.

ING has a 360 degree view on its customers, mainly within commercial banking, through a dashboard named as Vantage. GRID's integration into Vantage allows everyone at ING to quickly access parties' websites, view the legal and economic hierarchy of a party, identify key ING Relationship Managers, and see external ratings, for example, from Fitch, Moody's, or Standard and Poor's (Figure 2.1). Information from ING units, ING applications, and several external sources on millions of parties is also immediately available to all ING employees. Furthermore, Vantage ensures the processes in which information is used comply with international regulations.

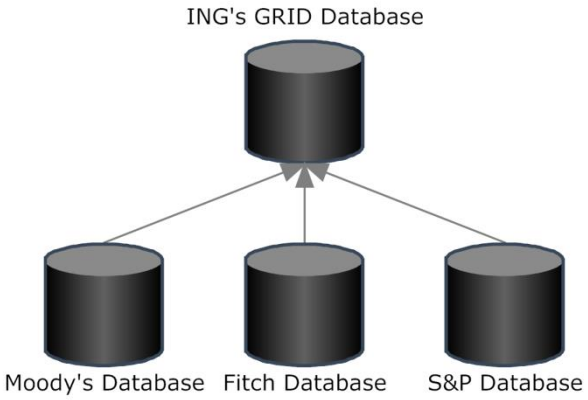


Figure 2.1 - Database schema

In GRID the data are organized into three levels: Data Groups, Data Categories and Data Fields [Raats 2013]. At the highest level, the data is organized into groups based on the way it was obtained in GRID. A *data group* is a set of data categories with common characteristics or features. All the data in a specific Data Group is treated and set-up in the same way in GRID. The following four different groups of data are distinguished:

1. Party Registration Data. It is the set of basic information should be registered for a Party in GRID;
2. Process Result Data. It is the set of required information that allows entering into an arrangement with a Party in GRID;
3. External Source Data. It is the set of information that is delivered to GRID by other sources (internal and external);
4. Referential Data. It is the set of references that connects data internally in GRID as well as with external sources.

A *data category* is a set of related data fields that is ordered together to make it easy for the user to find the information he is looking for. Data categories describe for example a Party characteristic like an address, a process result, a data set delivered by a specific source or a relationship between Parties. The data fields make up the lowest level of the structure holding the actual information. This information can either be:

- Free text;
- Number;
- Date;
- GRID ID;
- User ID;
- A value from a pre-defined list.

The main challenge in integrating risk rating data from the risk rating agencies with GRID is the development of a method that automates the reconciliation process and ensures the quality of the data.

2.5 Conclusion

Nowadays many organizations or governments are developing and selling securities. Such organizations are called issuers. In order to measure the creditworthiness of an issuer credit risk ratings are developed. These ratings are calculated by risk rating agencies. The three biggest ones are Moody's, Fitch, and Standard and Poor's. These agencies are selling the ratings to organizations that want to manage their risk exposure. With this information they can evaluate the issuers that they conduct business with. Important clients of the agencies are banks, who want to evaluate if a potential business customer of the bank will be able to pay its liabilities. After the economic crisis, according to Basel II, banks should conduct more rigorous credit analysis by using their own credit risk rating system in order to manage properly the externally rated securitization exposure. Even though they have their own internal rating system, they still buy risk ratings from the external risk rating agencies. The rationale behind this decision is the desire to reduce the risk to a minimum level, since doing business with a client that cannot pay its liabilities will have a severe effect on the profitability and the reputation of the bank.

ING has its own internal risk rating system. These risk ratings are the primarily source for evaluating and monitoring ING's customers. Nevertheless, it also buys risk rating from the three biggest risk rating agencies: Fitch, Moody's and Standard and Poor's. Account managers, credit analysts and risk managers are using the CCRM Credit Risk Systems Portal to manage and monitor the transaction approval process. In this intranet based collection of modules, the user can retrieve issue and issuer related financial and market data. The data is stored in the Global Relationship Identifier Database. GRID is a global relational Sybase database that contains approximately one million people, one million contacts and eight million legal entities.

3 Data reconciliation

3.1 Data reconciliation

Nowadays the importance of data is increasing. Both the scientific and organizational sector are developing techniques and methods to handle data better in order to retrieve as much information and knowledge as possible. Data is an elementary description of things and can be categorized as internal or external and as structured and unstructured. For example, data that are included in XML files are structured, while data that are included in word documents are unstructured. On the other hand, information is organized data that has a meaning and is valuable. When data or information is used in a business decision process, then it is known as knowledge [Caron 2014]. For example, a list of scores of some students in mathematic is considered as data. The average score of all the students for this course is information. If the teacher knows that all students that have scored lower that the average should work harder, then this is knowledge. Taking into account the increase of data volumes and the Big Data trend [Caruso 2000], the necessity for better manipulation of data is crucial.

Organizations retrieve data from different internal and external sources. These data sources can be homogeneous or heterogeneous. The main challenge is how they can retrieve the data in a unified view, while maintaining the quality of information, such as timeliness, accuracy and completeness [Inghaln et al. 1999]. This phenomenon is known as *data integration* and is part of the Business Intelligence framework. Data integration is needed at both schema and instance level [Lawrence et al. 2002; Sattler et al. 2003]. There are two main techniques for integrating heterogeneous data sources. The first one is data warehousing. Data warehousing is using the Extract, Transform and Load (ETL) procedure to transform data in a way that is compatible with each other and therefore be shown in a unified way. Data is extracted from each source, afterwards it is transformed in an appropriate form and in the end it is loaded to a front-end business intelligence system. The second way is developing a mediated schema in order to retrieve data directly from the original data source. There are two mediated approaches based on the mappings between the data sources and the global schema [Wiederhold 1992]:

- Global-as-view (GAV): The global schema is required to be expressed as a view on the data sources [Lenzerini 2002; Bouzeghoub et al. 2002]. According to Lenzerini (2002), this approach is effective when the data sources are stable
- Local-as-view (LAV): This approach requires that the global schema is independent to the data sources and every source is defined as a view over the global schema [Lenzerini 2002]. This approach according to Lenzerini (2002) processes the queries easier as it takes advantage of the mapping and as it is based on a simple unfolding strategy.

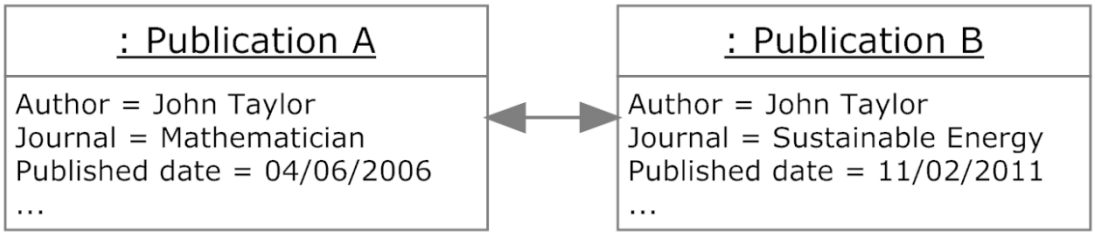
One important aspect of data integration is *data reconciliation*. Data reconciliation ensures the consistency of the data and is used when the external data must be matched with the internal data. Furthermore, data reconciliation is used to identify duplicate records within an internal database. A number of different definitions exist for the concept of data reconciliation. Carusso et al. (2000) describe a data reconciliation and data quality tool that uses a number of pre-processing and matching rules to identify and remove the duplicates in a database. According to Crowe (1996), data reconciliation is "the procedure of optimally adjusting measured data so that the adjusted values obey the conservation laws and other constraints". Data reconciliation has been observed in various sectors. Spindler (2014) proposes a data reconciliation technique to remove redundant equations for a given plant layout in wastewater treatment systems, while Özyurt et al. (2004) describe the importance of error detection procedures through data reconciliation to reduce the effect of gross errors in chemical processes. Gross errors are very important errors that cannot be characterized as random or systematic. For the identification of the gross errors, various methods have been used, such as correntropy estimators [Chen et al. 2013] After reviewing many scientific papers, we concluded that *data reconciliation*, in the context of business intelligence, deals with the decision on whether data descriptions from different sources refer to the same real world entity [Lenzerini 2002] and is considered as a part of data integration [Bizer 2012], which involves the provision of a unique view of the combined data.

In other words, data reconciliation deals with the identification of two different data objects that represent the same real world object. This problem has been found as entity matching [Ioannou et al. 2013], entity linkage [Ioannou et al. 2008], entity resolution [Singhal et al.], object matching [Doan et al. 2013], object reconciliation [Noessner et al. 2010], record linkage [Ioannou et al. 2013], merge-purge [Caruso 2000], deduplication [Caruso 2000], entity identification [Ioannou et al. 2013] and reference reconciliation [Dong et al. 2005]. Reference reconciliation deals with the problem of identifying when different set of attributes within a dataset is related to the same real world object. For example, Dong et al. (2005) have developed an algorithm that exploits the context information, propagates reconciliation and enriches the references in order to correctly match the references. On the other hand, F. Sais et al. (2007) use a logical method L2R that provides reconciliation with 100% precision.

When organizational data are reconciled, one of the most common techniques that is used by the organizational and scientific community is company name matching. This technique uses the names of the companies to identify and match the objects. First of all, the data are cleaned and transformed into a standard form to facilitate the reconciliation. Afterwards, some matching methods are applied. This technique is also widely used in patent analysis and harmonization. Magerman et al. (2006), harmonize the names through two stages. The first stage consists of data pre-processing, such as character and punctuation cleaning. The second stage consists of name cleaning. The common company words are removed, the abbreviations are translated and the spelling variations and the umlauts are harmonized. The focus on this approach is on the maximization of the accuracy. For further development, they suggest the usage of string matching algorithms, automatic acronym generations and the introduction of address information in order to maximize the accuracy and the number of matched entities. Thom et al., after preprocessing and standardizing the data, they use

two approaches for matching the companies' names. The first one is called dictionary-based approach and it uses a collection of large datasets of names and name variants to match the entities, while the second one, which is called rule-based approach, builds a set of rules for identifying similarities between names. More specifically, they use edit distance and token-based distance algorithms. Peeters et al. (2010), for harmonization of the patentee names, they focus only on the name similarity. After pre-cleaning the data, they are defining search keys and selection of new harmonized names and afterwards they use approximate string matching algorithms to match them. According to Magnani et al. (2007), the company name matching is a subcategory of string matching. The first stage of the reconciliation process is data preparation and analysis. The second stage is data harmonization, which transforms the data for the matching stage, which is the final stage of the reconciliation process.

Apart from the business community, the scientific community is also concerned with this topic. For example, in bibliometrics, where it is important to identify correctly the authors of published scientific papers in several databases. The problem is illustrated with the following question. Is John Taylor, who is the author of Article 1, the same person with the author of Article 2 with the name John Taylor? This importance of this problem is very significant as the author in a publication is much more than a string in a database. The publication is a mental property of an author and should not be assigned to a different person.



Picture 3.1 - Example of reconciliation problem in a bibliographic database

The data reconciliation process can be divided into three main categories:

1. Data selection;
2. Data cleansing;
3. Matching.

3.1.1 Data selection

At this stage, the attributes of the records that are going to be used for matching are being selected. There are many criteria that should be taken into account during the selection procedure. First of all, the attributes should represent a unique characteristic of the object, which will assist in the matching procedure. For example, a useful attribute for reconciling authors is the date of birth, which in combination with the full name of the author can indicate whether an "author A" is the same person with "author B". Moreover, using the gender to reconcile them probably will not give any useful information. Another characteristic that should be taken into consideration is the quality of the data of an attribute, such as the correctness of the data or missing values. For example, if the attribute

data of birth has too many empty values, then it should not be used in the reconciliation process. If the data is “dirty”, for example it contains errors, some techniques can be applied to improve its quality and then be used in the reconciliation process.

3.1.2 Data cleansing

Data cleansing or data cleaning or scrubbing is used to identify and remove or correct the errors and the inconsistencies of data in order to improve its quality [Naumann et al. 1999]. This process is mainly done automatically, but sometimes manual effort is required due to inconsistent anomalies. According to Maimon et al. (2005), the data cleansing process is divided into three sub-processes:

- Definition and determination of error types;
- Search and identification of error instances;
- Correction of the uncovered errors.

The first two are usually semi-automated processes that focus on identifying and correcting the errors, while the third one is mainly done manually in order to correct the errors that could not be identified. For the error detection several methods have been applied in the literature according to the type of the data. Some of these methods work on categorical data, while other work on quantitative data. Popular ones are:

- Pattern matching: This method uses different patterns for the identification of the records that have similar characteristics [Naumann et al. 1999]. The ones that do not have similar characteristics are called outliers and should be cleaned;
- Clustering methods: This method uses, for example, Euclidian distance to identify the outliers; [Naumann et al. 1999];
- Association Rules: This method uses some rules to identify the records that have similar characteristics. If a record does not follow these rules, then it is considered as an outlier [Naumann et al. 1999];
- Statistical methods: This method uses statistics, such as mean or standard deviation, to identify the outliers [Naumann et al. 1999; Singhal et al.];
- String parsing: This method is performed for the detection of the syntax errors by analyzing a string of symbols [Singhal et al.]. The string characters are split into tokens, analyzed and then the tokens are formatted in a structured way.

Apart from these methods, some basic transformational rules are applied. Some rules focus on changing the data from its original format to the format that is expected in order to harmonize the reconciliation procedure. When this is applied on the instance level, it is known as standardization or normalization. Two widely used transformational rules are punctuation standardization and company name standardization [Magnani et al. 2007]. The first one removes the punctuation marks in a string, while the second one translates the abbreviations for company names. For example, “Co” is translated to “Company” in order to ensure that the data use the same abbreviations. Some other rules focus on multilingual translation (translate ‘Municipio’ to ‘Municipality’) and some others on character translation (translate ‘ë’ to ‘e’). All these rules are often solved through SQL queries.

3.1.3 Matching

At this stage, matching rules are used to match the values of different attributes. Usually, a combination of attributes is used for the reconciliation in order to maximize the accuracy of the matching procedure. Some techniques focus on the identification of records in a data source, that match exactly or approximately with some attributes, using some string matching algorithm. Other techniques filter out records that have some specific characteristics, such as empty values in attributes or classify mismatched records based on a measure [Naumann et al. 1999].

3.2 Data reconciliation related to risk rating data

In the existing literature, there is little focus on data reconciliation related to risk rating data. In our opinion, there are two main reasons. First of all, risk rating data are confidential as companies do want to share it with its competitors and many organizations are reluctant to share their methods on how they reconcile it. In addition, risk ratings are not provided for free and the cost of purchasing them is high. Thus, there are few publications on this topic. The second reason is related to the type of data. As mentioned in Chapter 2, the attributes that can be used in the reconciliation process are often not sufficient to accurately match the issuers. Generally, the unique identifier, the legal name and the country of residence of a record is provided by the risk rating agencies, which does not give enough freedom or information to the researchers to develop a prototypical method. However, there is related literature on the standardization of company and organization names. For example, in patent analysis. Furthermore, risk rating data is often not clean and most of the researchers make assumptions on the correctness of the data. For researchers, such data is often not directly available. Consequently, they prefer to apply their methods and techniques on different types of data.

On average, the legal names of the issuers consist of long strings (on average 27 characters) and there are few variations on the legal names as organizations use the official registered names of the issuers. Therefore, the majority of the variations are due to mistakes during the manual insertion into the database or due to encoding transformations. In addition, the difference between the countries of residence of two issuers that are indeed the same, is frequently due to the confusion with the country of incorporation. The country of residence or country of domicile is the country where the company has its permanent address, while the country of incorporation is the country where the company is legally registered.

3.3 Challenges of data reconciliation

In this section, the main challenges of the data reconciliation process are identified. According to Ioannou et al. (2013), there are five categories of variations that set hurdles to the reconciliation process [Ioannou et al. 2013; Müller et al. 2005]. For each category, there are some sub-categories, which make the problem more specific. Apart from these five categories, we identify an additional one, which is “coverage anomalies” [Singhal et al.]. The categories:

1. *Syntactic variations*: This category contains differences in the values of the attributes that are used for comparison. There are five sub-categories:
    - Misspellings: The value of the attribute may not be spelled correctly. For example, “ING BV Incorporation” vs. “ING BV Incropration”.
    - Homonymity: The value of the attribute, usually a name or an address, can be the same for two objects, even though they are different. For example, “Wall Street” in London vs. “Wall Street” in New York.
    - Different order: The order of the words in an attribute can be different. For example, “Dimitrios Routsis” vs. “Routsis Dimitrios”.
    - Different standards: The way that a value, e.g address, is written can vary between organizations or countries. For example, “Favierou 37 Street” vs. “37-Str Favierou”.
    - Abbreviations: The value of the attribute may contain an abbreviation for a word. For example, “ING Inc.” vs. “ING Incorporation”

2. *Structural variations*: This category contains differences in the structure of the attributes that are used for comparison.
    - Different number of attributes: For the description of a characteristic of an object, a data source may use one attribute, while another data source may use a set of attributes. For example, the data source A uses the attribute “full name” to describe the full name of issuer, while the data source B uses the attributes “First name” and “Last name” to describe it.

3. *Semantic variations*: This category contains cases where the value of the attributes is the same, but the objects are different in reality. There are two sub-categories:
    - Synonyms: Synonyms derives from the Greek word «Συνώνυμο» and means the same name. For example, “Rich” vs. “Wealthy”.
    - Multilingualism: The values of the attributes contain words that have the same meaning but are written in different languages. For example, “Municipio of Torino” vs. “Municipality of Torino”.

4. *Evolution of attributes*: The values of the attributes of an entity do not remain stable over the time. For example, an issuer may have moved to a different address and one data source may have not updated its records with the new address.

5. *Association network variance*: Some objects in a data source may be connected to each other in a way that they are forming a network. Consequently, in order to identify and thus use them for the reconciliation procedure, this network must be defined and its associations must be analyzed.

6. *Coverage anomalies*: This category contains cases where some information that describes the object is missing. Two sub-categories are:
    - Missing attributes: The provided attributes for a specific object may be too inadequate for the reconciliation process. For example, data source A provides only the "Country of residence" of an issuer. It is impossible to reconcile the issuer by using only the country of residence as they will be too many matches.
    - Missing values: The values of an attribute can be missing, and thus the comparison using this attribute is not applicable.

3.4 Approximate string matching algorithms

The majority of the attributes that are used in the reconciliation process are of the string data type. For the comparison of strings several techniques are being used, including string matching algorithms. There are two kinds of string matching algorithms:

- Exact string matching algorithms;
- Approximate string matching algorithms.

The first category returns the records that match exactly. Charras et al. (2004) have written a handbook that contains all the exact string matching algorithms. Important exact string matching algorithms are:

- Brute force algorithm;
- Knuth-Morris-Pratt algorithm;
- Boyer-Moore algorithm.

The second category returns the percentage of match between two strings. In this section, we are focusing on approximate string matching algorithms and after explaining the most popular ones, we are evaluating them. Important approximate matching algorithms:

- Levenshtein distance algorithm [Levenshtein 1966];
- Jaro-Distance algorithm [Jaro 1989];
- Jaro-Winkler algorithm [Winkler 1990].

3.4.1 Levenshtein distance algorithm

Levenshtein distance is an edit distance algorithm that according to the National Institute of Standards and Technology returns "the smallest number of insertions, deletions and substitutions required to change one string or tree into another"[Black 2014]. The complexity of the algorithm is $\Theta(m \times n)$, where m is the length of the first string and n the length of the second string.

The algorithm uses a matrix (*m* x *n* dimensions) to calculate the distance. The first column is initialized from zero to m and the first row from zero to n. Then, it checks each character from i equals 1 to m and from j equals 1 to n and sets the cell[I,j] equals to the minimum of the following equations:

- Cell[i-1,j] + 1;
- Cell[I,j-1] + 1;
- Cell[i-1,j-1] + cost, where cost equals to zero when the character i of the first string is equal to the character j of the second string, and equals to 1 where the character i of the first string is not equal to the character j of the second string.

After filling in all the values in the matrix, the distance between the two strings is the value of the cell [m,n]. For the calculation of the percentage of match between two strings, the following equation is used:

$$Percentage = \left[1 - \frac{Levenshtein\ Distance}{max\{length(String1), length(String2)\}}\right] * 100 \qquad \text{Eq. 1}$$

In Table 3.1, we calculate the Levenshtein distance between “ING BANK INC” and “ING BANKO INC”.

		I	**N**	**G**		**B**	**A**	**N**	**K**		**I**	**N**	**C**
	0	1	2	3	4	5	6	7	8	9	10	11	12
I	1	0	1	2	3	4	5	6	7	8	9	10	11
N	2	1	0	1	2	3	4	5	6	7	8	9	10
G	3	2	1	0	1	2	3	4	5	6	7	8	9
	4	3	2	1	0	1	2	3	4	5	6	7	8
B	5	4	3	2	1	0	1	2	3	4	5	6	7
A	6	5	4	3	2	1	0	1	2	3	4	5	6
N	7	6	5	4	3	2	1	0	1	2	3	4	5
K	8	7	6	5	4	3	2	1	0	1	2	3	4
O	9	8	7	6	5	4	3	2	1	1	2	3	4
	10	9	8	7	6	5	4	3	2	1	2	3	4
I	11	10	9	8	7	6	5	4	3	2	1	2	3
N	12	11	10	9	8	7	6	5	4	3	2	1	2
C	13	12	11	10	9	8	7	6	5	4	3	2	**1**

Table 3.1 - Example of Levenshtein distance calculation

The Levenshtein distance of these strings is 1. After using the equation 1, we find that the percentage of match between these two strings is 92.31%.

3.4.2 Jaro-Distance algorithm

The Jaro-Distance algorithm an edit distance algorithm that according to the National Institute of Standards and Technology returns ”the weighted sum of percentage of matched characters from each file and transposed characters” [3.27]. This algorithm uses the following equation to calculate the string edit distance between two strings, string1 and string 2.

$$Jd = \begin{cases} 0, if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|string1|} + \frac{m}{|string2|} + \frac{m-t}{m}\right) \end{cases} \quad \text{Eq. 2}$$

In this equation, m is the number of matched characters. Matched characters of the two strings are the characters that are the same and their position is between the distances that are calculated with the following equation:

$$Distance \leq \left|\left(\frac{\max(string1, string2)}{2}\right) - 1\right| \quad \text{Eq. 3}$$

On the other hand, t is the number of transpositions divided by two.

In the following example, we calculate the Levenshtein distance between string1 = "ING BANK INC" and string2 = "ING BANKO INC". The number of matching characters (m) is 9. The number of transpositions is 2. After inserting the number in equation 1, we find that the Jaro Distance for these two strings is 0.8603.

3.4.3 Jaro-Winkler algorithm

The Jaro-Winkler algorithm an extended version of the Jaro-Distance algorithm that according to the National Institute of Standards and Technology "increases the Jaro-Distance measure for matching initial characters, then rescaled it by a piecewise function, whose intervals and weights depend on the type of string" [Black]. This algorithm uses the following formula to calculate the match:

$$Jw = Jd + l * p * (1 - Jd) \quad \text{Eq. 4}$$

Jd is the distance that is calculated with the Jaro-Distance, l is the length of the common (maximum four) prefixes between the two strings and p is constant that should not be higher than 0.25. Usually, it is set to 0.1.

After examining the previous algorithms, we conclude that Jaro-Distance and Jaro-Winkler are more suitable for matching names and short strings. More specifically, Jaro-Winkler gives better results, as we shall see in Chapter 5, if the name starts with the same characters. On the other hand, if there is a wrong character or a character is missing in the prefix of a string, Jaro-Winkler gives worse results. Thus, these two characters are better when the attribute that is compared contains one string and not a combination of strings. The Levenshtein distance algorithm has better behavior and gives better results when comparing long strings or a combination of strings. This is due to the fact that it can substitute a character, which according to the algorithm will cost only one action and not two. Nevertheless, this does not imply that its performance deteriorates when comparing short strings.

3.5 Conclusion

Data reconciliation, in the context of business intelligence, deals with the decision on whether data descriptions from different sources refer to the same real world entity [Lenzerini 2002]. It is considered to be part of data integration [Bizer 2012], which involves the provision of a unique view of the combined data. The main challenges of this process are caused by syntactic variations, structural variations, semantic variations, evolution of attributes, association network variances and coverage anomalies. Even though a plethora of diverse techniques and method are being used, the most common steps in this process are data selection, data cleansing and matching. Data selection deals with the selection of the attributes that will be used to the reconciliation process. Data cleansing cleans and transforms the data in a way the will facilitate the matching procedure. For matching the records, approximate string matching algorithms are being used that compare the strings of the attributes of the records. The widely used algorithms are Levenshtein distance, Jaro-Distance and Jaro Winkler.

4 Method for automated data reconciliation of risk ratings

4.1 Introduction

This chapter presents a method to automate the data reconciliation process of risk ratings. The main focus of this conceptual design is, first of all, the improvement of the data quality of the internal database by identifying only correct matches and then the minimization of the manual effort and the time of execution of the reconciliation process. Thus, even though the majority of the common challenges of data integration processes are applied in our case, data analysis is required in order to identify to what extent they exist and which of them should be resolved based on their frequency. For example, if the data is in the same language, then no linguistic transformation will be required.

4.2 Data analysis

For quantitative studies, sampling techniques for selecting the elements from which the information will be collected are very common. The main advantages of these techniques are time efficiency, high reliability and low cost. On the other hand, these methods have several disadvantages, such as chances for bias, problems of accuracy, inadequacy of samples and chances of committing errors in sampling [Ghauri et al. 2005]. Since, the populations of the monthly files were not too large and since high levels of accuracy were preferable, we decided to select one entire file from Moody's, Fitch and S&P for our data analysis and not a sample from each file. Due to the fact that these files contain all the available issuers of each risk rater at that period, the procedure of selecting the sample file was out of importance. Therefore, the simple random sampling technique was used for selecting one file for each risk rater.

For the preliminary study, the monthly files 20140504.txt, 20140501.txt and 20140701.txt from Moody's, Fitch and S&P accordingly were selected. The first step of the analysis was to identify and exclude all the issuers that could be matched on the cross-reference. If there was an active organization in GRID that has the same value with the external ID, then these records were matched. One of our main assumptions was that these cross-references were correct and point to the same issuer. After excluding these records, since the statutory legal name is unique for each organization in its country of residence and since all the risk raters provide this information in their files, we matched the remaining unmatched issuers with GRID organizations based on the statutory legal name, only if they matched 100% and without transforming the data prior to the matching. In Table 4.1, the results of the first two steps of our analysis can be seen. The majority (90.63% on average) of the issuers was matched on the cross-reference and few (2.2% on average) were matched on the legal name. On average 7.17% of the files could not be matched on either the cross-reference or the legal name.

	Moody's		Fitch		S&P	
File	20140504.txt		20140501.txt		20140701.txt	
Total Records	10.910	100%	7.411	100%	10.104	100%
Matched on XREF	9.373	85.91%	7.044	95.05%	9.189	90.94%
Unmatched on XREF	1.537	14.09%	367	4.95%	915	9.06%
Matched on Legal Name	335	3.07%	42	0.57%	298	2.95%
Unmatched on Legal Name	1.202	11.02%	325	4.39%	617	6.11%

Table 4.1 - Analysis of sample files

For the remaining unmatched issuers, in order to identify the reason why they did not match, we analyzed manually and categorized them accordingly. The manual check was conducted using GRID's search engine. The majority of the categories of mismatches were same for all the files, but the frequency was slightly different. During the analysis, it was observed that many issuers fell into more than one category. For example, 'State of Malta' vs. 'Malta, State of' belongs to the category of i) Abbreviations and ii) Different order. Nevertheless, it was assigned only to the category that fitted best and not to both. The Table 4.2 shows the categories of mismatches and their explanations, and the Table 4.3 shows the number of records for each file for each category. In the Moody's file, there were some issuers, whose legal name contained the word "(New)". This is an indication that the issuer is new at Moody's database and is only used for Moody's internal purposes. Therefore, these words were removed prior to the analysis of the data.

Status	Explanation	Example (GRID vs. External)
Does not exist	The issuer does not exist in GRID	-
Punctuation Marks	The punctuation marks differ	Sedgwick, INC. vs. Sedgwick INC
Abbreviations	There are abbreviations for some words in the legal name	CO vs. Company
Extra Description	The issuer's legal name has more words than GRID's one.	Oleoducto Central vs. Oleoducto Central SA
Inadequate Description	The issuer's legal name has insufficient description	Standard Oil Co Inc vs. Standard Oil Co
Complex	It's complex to be automatically be resolved	Siam Commercial Bank Public PCL vs. Siam Commercial Bank Public CO. LTD. (CI)
Wrong Abbreviation (?)	The abbreviation of the issuer is different from GRID's abbreviation	CommScope Holding Inc vs. CommScope Holding CO
Different Order	The legal name is the same, but in different order	The EW Scripps CO vs. EW Scripps Co The
Diacritics and Special Characters	There are special characters in the legal name, such as Ş, ü, ñ, í, ┴., ┌, ╓	Collectivit ¬S Territoriales vs. Collectivit os Territoriales
Space	There is an extra or more spaces in the legal name	Lochpe Maxion SA vs. IochpeMaxion SA
Different Language	The words are in different languages, but have the same meaning	Municipality of Patras vs. Municipio of Patras
Country Part of the L.Name	The country is part of the legal name	Enbridge Inc vs. Enbridge (U.S.) Inc

Municipality	The word municipality is used to describe the city	City of Tehuacan vs. Tehuacan, Municipality
Former Name	The former name of the issuer is part of the name	Wachovia Bank FSB vs. Wachovia Bank FSB (formerly World Savings Bank FSB Texas)

Table 4.2 - Categories of mismatches and explanations

Status	**Moody's**		**Fitch**		**S&P**	
	Records	*Percentage*	*Records*	*Percentage*	*Records*	*Percentage*
Does not exist	594	49.42%	236	72.62%	496	80.39%
Punctuation Marks	347	28.87%	40	12.31%	42	6.81%
Abbreviations	120	9.98%	15	4.62%	10	1.62%
Extra Description	41	3.41%	8	2.46%	9	1.46%
Inadequate Description	34	2.83%	17	5.23%	30	4.86%
Too Complex	30	2.5%	1	0.31%	9	1.46%
Wrong Abbreviation (?)	15	1.25%	1	0.31%	10	1.62%
Different Order	7	0.42%	2	0.62%	2	0.32%
Special Characters	5	0.25%	1	0.31%	6	0.97%
Space	3	0.17%	0	-	3	0.49%
Different Language	3	0.58%	2	0.62%	0	-
Country Part of the L.Name	2	0.25%	1	0.31%	0	-
Municipality	1	0.08%	0	-	0	-
Former Name	0	-	1	0.31%	0	-

Table 4.3 - Categories of mismatches for each credit rater

In Table 4.3, it can be clearly observed that the majority of the mismatches were due to the fact that:

i. The issuers did not exist in GRID;
ii. The issuers had different punctuation marks with the organizations in GRID;
iii. The issuers or the organizations in GRID had abbreviated words.

On average, these categories consist of 89% of the total population, which is a relatively large percentage. In Fitch's and S&P's file, the majority of the issuers did not exist in GRID, while in the Moody's file half of the issuers did not exist in GRID. During the analysis of the data, all the abbreviations that were met were stored in a translation table, which can be seen in Table 4.4. These abbreviations were used in our methodology, which is described in chapter 4.3.

Abbreviations	
Code	*Description*
PCL	PUBLIC COMPANY LIMITED
Ltd	Limited
Corp	Corporation
COMM	Commercial trust
SPA	Societa Per Azioni
CO	Company
Bhd	BERHAD
GOVT	Government
AG	Aktiengesellschaft
Holding	Holdings
New York	NY
PSP	Public Sector Pension
AS	Anonim Sirketi
Inc	INCORPORATED
TRS	Trust
Intl	International

Table 4.4 - Abbreviations and their description

4.3 Design of the method

Based on the results of the data analysis, we designed our method for automated data reconciliation with focus on the maximization of the data quality in the internal database and the minimization of the manual effort and the time of execution. Therefore, all the available sources were evaluated and only the ones that added significant value were incorporated. The method consists of the following five steps:

Step 1. Data Cleansing;
Step 2. Match on Cross-reference;
Step 3. Match on Legal Name;
Step 4. Match using BvD Cross-references (only for Moody's);
Step 5. Approximate String Matching Algorithm.

1st Step: Data Cleansing

In the first step, the majority of the data errors and inconsistencies are eliminated in order to facilitate the matching process on the legal name and country of residence. The process of cleaning the legal name consists of three steps, 2 of which are common for all the risk raters and one which is only used in Moody's file.

- The first step is only used for Moody's and is the elimination of the word "(New)" that exists in the legal name. As it was mentioned in the data analysis chapter, this word appears only in Moody's file and not in S&P's and Fitch's. This is an indication that this issuer is new to Moody's database and is used only for internal purposes.
- The second step is the elimination of all the punctuation marks (Table 4.5) for both the internal and external file.

- The third step is the translation of the abbreviations that are stored in the translation table (Table 4.4). Both the GRID and the external file are cleaned, because it was observed that there were inconsistencies even in the same database. For example, in GRID one organization contained the word Incorporated in its legal name and another organization the abbreviation Ltd.

	Characters
Punctuation Marks	. , ( ) ! # @ $ “ ” % & ‘ ’ + - * = / ; : > < [ ] ? { } | \ ^ ~

Table 4.5 - Punctuation Marks

These categories cover approximately 90% of the population. Even though we could have resolved more categories of mismatches at this step, we decided not to clean them for two main reasons. First of all, the only available database management system (DBMS) in ING was Microsoft Access 2010[5]. Microsoft Access is a user friendly Structured Query Language (SQL) relational database that is intended to handle small amount of data. Unfortunately, for managing large amount of data, such as GRID that contain approximately 9.3 million records, it is not suitable as it has several storage (2 GB) and processing limitations. For example, translating the special characters in both GRID and external file takes more than 3 hours. Therefore, since this category of mismatches is not very common, the time of execution outweighs the added value from cleaning. Secondly, most of the records (10%) will be resolved and matched by the implementation of the string matching algorithm in Step 5. The string matching algorithm is implemented in Python[6] language in Portable Python 2.7.6.1[7] environment. Python is a programming language that gives the user the possibility to work quickly and in a powerful way. Thus, resolving these categories using the String Matching Algorithm in Python is much faster than transforming the data in Microsoft Access.

The process of cleaning the country of residence consists of the translation of the country codes. The countries of residence are displayed through ISO 3166 codes[8] (Table 4.6). ISO 3166 is the International Standard for the country codes and the codes for their subdivisions. The country codes can be represented either as a two-letter code (alpha-2), a three-letter code (alpha-3) or a three digit numeric code (numeric-3). S&P is using ISO3 codes and Fitch is using a full description of the country, while in GRID the countries of residence are displayed in two ways. The first one is using the ISO2 codes and the second one is using the full description of the country. For the comparison of the countries, Fitch’s and S&P’s countries of residence will be transformed to ISO2 codes. The decision on transforming Fitch and S&P was based on the fact that their files contain significantly less records than GRID and the transformation will take less time. In addition, we decided to transform them into ISO2 codes, because they contain only 2 characters and the comparison will be quicker executed.

[5] http://office.microsoft.com/en-001/access/
[6] https://www.python.org/
[7] http://portablepython.com/
[8] http://www.iso.org/iso/country_codes.htm

ISO3 Code	ISO2 Code	Description
EGY	EG	Egypt
ECU	EC	Ecuador
DZA	DZ	Algeria
DOM	DO	Dominican Republic
DMA	DM	Dominica
DNK	DK	Denmark
DJI	DJ	Djibouti
DEU	DE	Germany
CZE	CZ	Czech Republic
CYP	CY	Cyprus
…	…	…

Table 4.6 - Translation table of the Country of Residence

2nd Step: Match on Cross-Reference

In this step, the issuers are matched based on the existing cross-references in GRID. If the cross-reference of a record in GRID has the same value with the ID of the issuer in the external file, then these are matched. As it was mentioned in chapter 4.2, the main assumption is that these cross-references are valid and point to the correct issuer. If not, then the database should be cleaned. All the matched issuers are stored in the file "Matched_ID". The remaining unmatched ones are stored in the file "Unmatched_ID".

For example, the GRID IDs '36012333' and '36012435' in Figure 4.1 match with Fitch's IDs '80088977' and '80640638', because GRID's *Fitch XREF* attribute has the same value with Fitch's *ID*. In other words, these records are linked due to the existing cross-references. On the other hand, Fitch's IDs '80090783' and '91293490' do not match, because there is no cross-reference in GRID that point to these records.

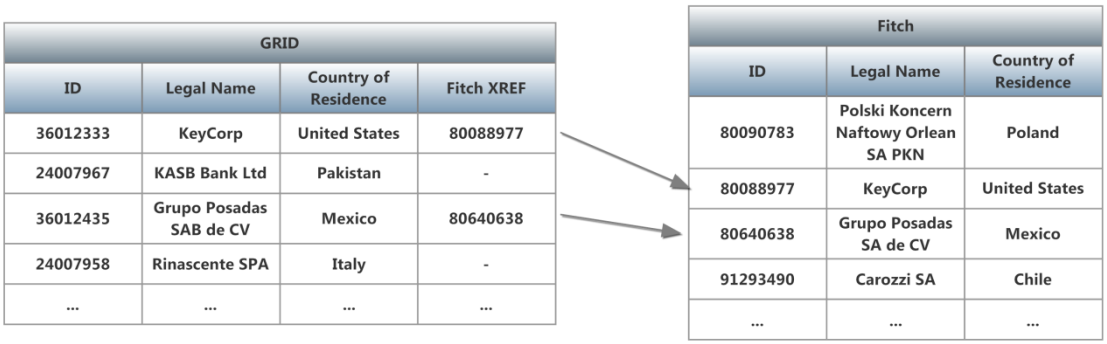
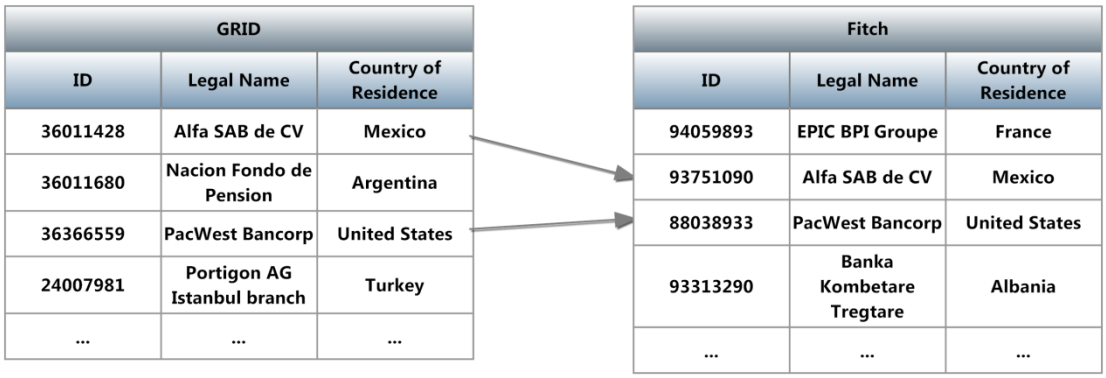


Figure 4.1 - Example of matching Fitch with GRID based on the cross-reference

3rd Step: Match on Legal Name

After transforming and cleaning the data, the external issuers (Unmatched_ID) are matched with the internal organizations on the statutory legal name. If the external legal name is equal to the internal legal name (only 100% match), then they match and are stored in the file "Matched_LN". The remaining unmatched ones are stored in the file "Unmatched_LN". In the matching process for 100% match, the country of residence was not included because the legal name gives much more confidence that the country of residence and because it was observed that in some cases the legal name of an issuer matched correctly 100% with an organization in GRID, but the countries of residence were incorrectly not the same. In other words, the country of residence was often incorrect either in GRID or in the issuer file.



GRID		
ID	Legal Name	Country of Residence
36011428	Alfa SAB de CV	Mexico
36011680	Nacion Fondo de Pension	Argentina
36366559	PacWest Bancorp	United States
24007981	Portigon AG Istanbul branch	Turkey
...	...	...

Fitch		
ID	Legal Name	Country of Residence
94059893	EPIC BPI Groupe	France
93751090	Alfa SAB de CV	Mexico
88038933	PacWest Bancorp	United States
93313290	Banka Kombetare Tregtare	Albania
...	...	...

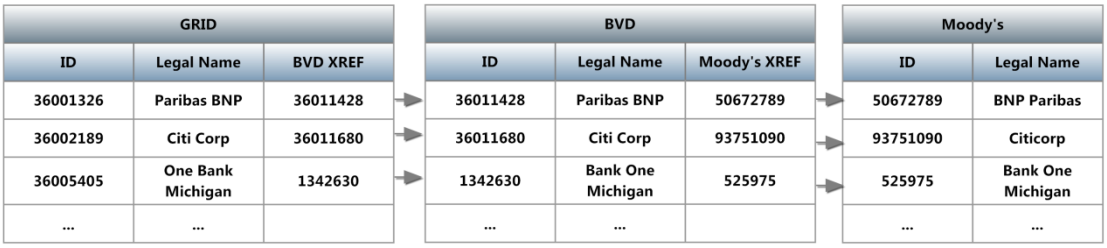
Figure 4.2 – Example of matching Fitch with GRID based on the Legal Name

4th Step: Match using BvD cross-references (only for Moody's)

ING buys data from Bureau van Dijk (BvD)[9] in order to enrich and update its database. Moody's is also client of BvD and as a result BvD has cross-references to link its records to the Moody's records. Due to the fact that BvD was willing to provide ING the cross-references, we used them to match more issuers. Unfortunately, BvD could only provide us with Moody's cross-references and not Fitch and S&P as well. Thus, this step is only used for Moody's. For every record where ING's BvD cross-reference is equal to BvD ID and BvD's Moody's cross-reference is equal to Moody's ID, the Moody's ID is matched with ING's ID. The matched records are stored in the file "Matched BvD" and the remaining unmatched ones are stored in the file "Unmatched BvD".

[9] http://www.bvdinfo.com/nl-nl/home

For example, in Figure 4.3 the Moody's issuer "50672789" was correctly matched with the GRID organization "36001326" using the BvD record "3611428".



GRID		
ID	Legal Name	BVD XREF
36001326	Paribas BNP	36011428
36002189	Citi Corp	36011680
36005405	One Bank Michigan	1342630
...	...	

BVD		
ID	Legal Name	Moody's XREF
36011428	Paribas BNP	50672789
36011680	Citi Corp	93751090
1342630	Bank One Michigan	525975
...	...	

Moody's	
ID	Legal Name
50672789	BNP Paribas
93751090	Citicorp
525975	Bank One Michigan
...	...

Figure 4.3 - Example of matching Moody's with GRID based on the BvD's Cross-references

5th Step: Approximate String Matching Algorithm

For the remaining unmatched issuers, an approximate string matching algorithm is applied on the statutory legal name. The approximate string matching algorithm finds the strings that match a pattern approximately and not exactly. The algorithm compares the external issuers with GRID organizations and returns the GRID organization with the highest match for each issuer. In defining the matching ratio, the country of residence is also taken into account, when is available, and the matching ratio is reduced whenever the country of residence of the issuer is different to the country of residence of the compared GRID organization.

The main objectives of this step are:

i. To correctly match as many issuers as possible to GRID's organizations;
ii. Identify and exclude the issuers that do not exist in GRID in order to reduce the manual reconciliation effort;
iii. Give accurate suggestions to the user for the issuers that should be manually reconciled to facilitate the matching procedure.

Thus, we identified two thresholds, one lower (Θ1) and one upper (Θ2), and based on the matching ratio, we categorized the issuers into three categories. Each category was stored into a different file. The values of the thresholds are calculated in Chapter 5.

Therefore, the output of this step is three files:

1. Unmatched (New): This file contains all the issuers that do not exist in GRID. All the issuers, whose highest matching ratio is below the lower threshold, are stored in this file.
2. Unmatched (Manual): This file contains all the issuers that may or may not exist in GRID. All the issuers, whose highest matching ratio is between the lower and the upper threshold, are stored in this file. These issuers should be manually reconciled, because the algorithm cannot 100% indicate whether it is a new issuer or it exists in GRID. This decision was based on one of our main goals, which was the improvement of the data quality in the internal database. If a wrong cross-reference is uploaded to the system, the identification and cleaning of this record requires double effort than the manual reconciliation of it. With the aim of facilitating the manual reconciliation, the three highest matching ratios /organization are stored in the file. In that way, there is a high probability that if the issuer exists in GRID, it will be one of these three records. For every record, the issuer's ID, issuer's legal name, GRID's legal name, GRID's ID, issuer's country of residence, GRID's country of residence and the matching score are shown to the user. As a consequence, the user can effortlessly identify the correct one.
3. Matched: This file contains all the issuers that are matched with GRID's organizations. All the issuers, whose highest matching ratio is higher than the upper threshold, are stored in this file. For every match, a new cross-reference is created and stored. All the new cross-references should be uploaded to GRID in order to improve the reconciliation procedure of the next incoming file of issuers.

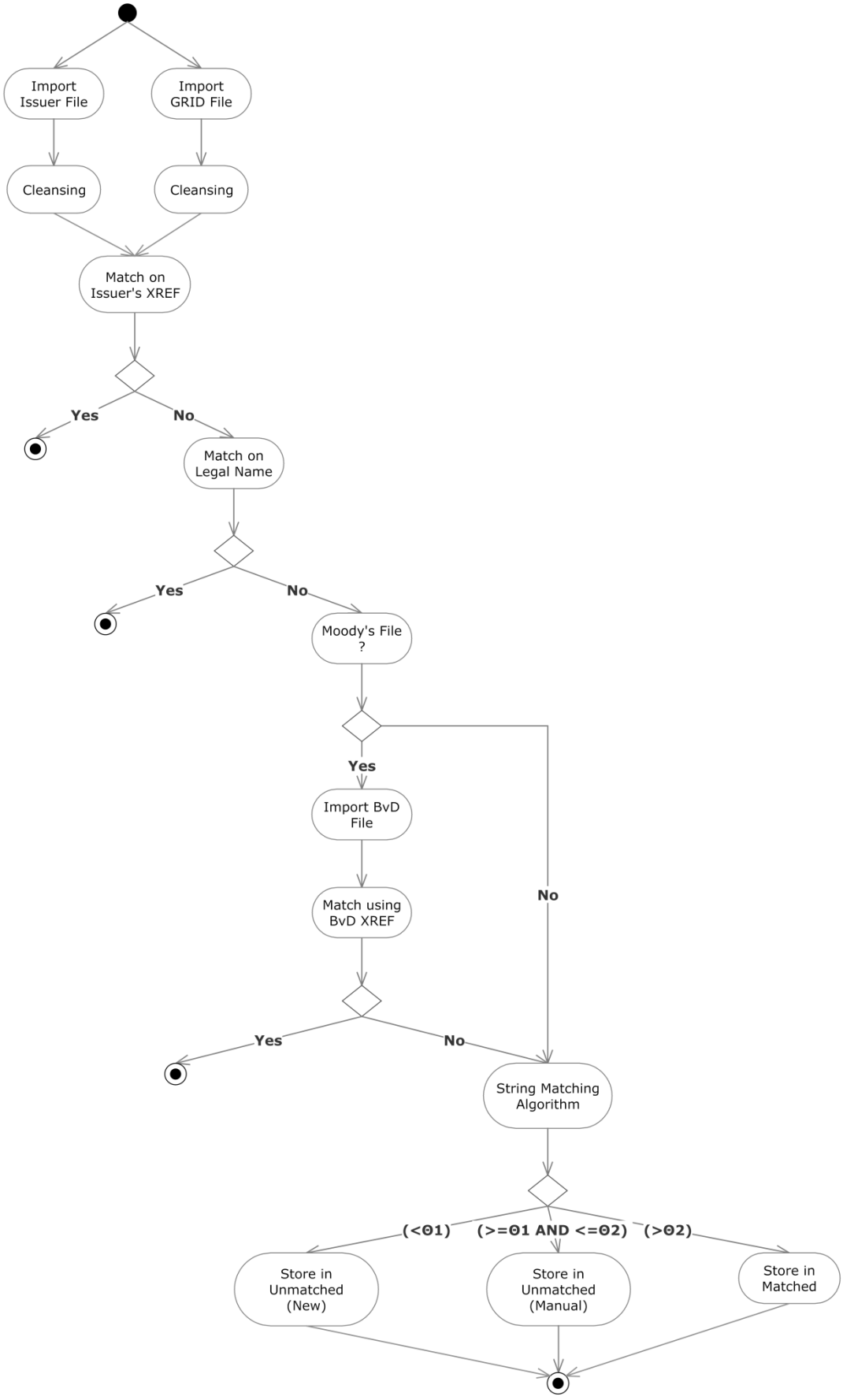


Figure 4.4 - UML diagram of the conceptual design of the method

4.4 Alternative configurations

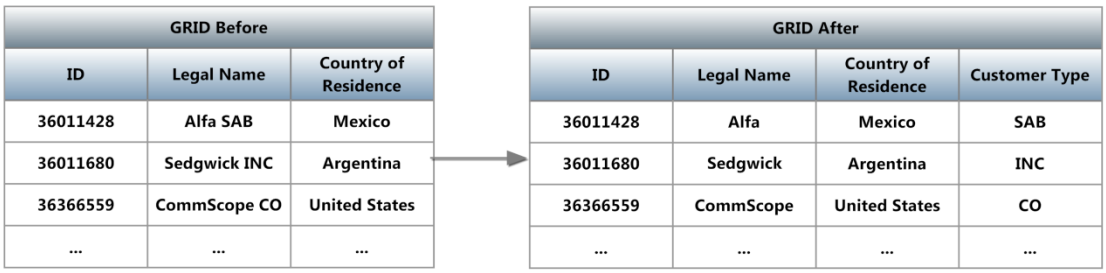
In this section, we propose some alternative configurations of our method. These configurations were not included in our method due to the quality of our data and the available infrastructure. In addition, the accuracy of our method was high and these configurations could not increase it any further. Nevertheless, they can add significant value to the reconciliation process for a different data set by reducing the processing time and increasing the number of matched entities. Two prerequisites for using them are:

- The data should have good quality;
- The infrastructure should enable quick processing of queries.

4.4.1 Customer type

Fitch and S&P did not provide the customer type of the issuer. Nevertheless, it was observed that the customer type was included in the statutory legal name of the issuer. For example, the statutory legal name of Sedwick is "Sedgwick INC", which contains the abbreviations INC that indicates that the issuer's type is incorporation.

In order to take advantage of the customer type, the abbreviations should be extracted from the legal name and inserted in a new attribute, called *customer type* (Picture 4.1). This should be done for both the internal and external data set. Afterwards, the issuer should be only compared with the internal data that have the same customer type. For example, if the type of the issuer is "Incorporation", then it should be compared only with GRID's records, whose customer type is "incorporation".



GRID Before		
ID	Legal Name	Country of Residence
36011428	Alfa SAB	Mexico
36011680	Sedgwick INC	Argentina
36366559	CommScope CO	United States
...	...	...

GRID After			
ID	Legal Name	Country of Residence	Customer Type
36011428	Alfa	Mexico	SAB
36011680	Sedgwick	Argentina	INC
36366559	CommScope	United States	CO
...	...	...	...

Picture 4.1 - Customer type configuration

GRID	Issuer
Financiere Gaillon 8 SA	Financiere Gaillon 8 SAS
Servicios Corporativos Javer SAPI de CV	Servicios Corporativos Javer SA PI de CV
Corporacion Azucarera del Peru S A	Corporacion Azucarera del Peru SA
Infraestructura Energetica NOVA SA de CV	Infraestructura Energetica NOVA SAB de CV
Alpek SAB de CV	Alpek SA De CV
Rottapharm	Rottapharm SPA
Azerenerji JSC	Azerenerji PJSC
Spar und Darlenhnkasse eG	Spar und DarlenhnkasseeG
SK Broadband CO Ltd	SK Broadband COLtd
PT BFI Finance Indonesia Tbk	BFI Finance Indonesia Tbk PT
PT Surya Artha Nusantara Finance	Surya Artha Nusantara Finance PT

Table 4.7 - Cases of incorrect customer types

This configuration was not part of our method, because the quality of the data was not very good and therefore the customer type could not be extracted correctly from the statutory legal name. In addition, in many cases (Table 4.7) the customer type was not correct. Consequently, if we had used this configuration, the accuracy of our method would be lower.

4.4.2 Clusters

This configuration identifies the name variants and clusters the issuers accordingly. The variations could be in spelling or in the way they appear within the database. Every cluster is stored in a different table. For example, “BSH Bosch und Siemens Aktiengesellschaft” and “BSH Bosch und Siemens Aktingeseelschaft” will belong to the same cluster.

Whenever an issuer is going to be reconciled, the method will first check if it belongs to an existing cluster. If it exists, it will check only the records of that cluster to identify if the issuer is the same with one of these records. After reconciling it, the cluster should be updated with the new issuer. If the issuer does not belong to any cluster, then it should be reconciled to the entire database. The main advantage of this configuration is that the processing time will be significantly reduced, as the issuer may not be compared with the entire database.

This configuration was not part of our method, because the available infrastructure had limitations on the speed of execution. Furthermore, after clustering our data, the outcome was clusters of a pair of records. Consequently, it did not add any significant value in our case.

4.5 Conclusion

This method consists of four common steps and one additional for Moody’s. At the first step, the issuers are matched on the existing cross-references. It is assumed that the cross-references are valid and point to the correct issuer. At the second step, both internal and external data is cleaned and transformed. The attributes that are cleaned are the statutory legal name and the country of residence. After cleaning the data, at the third step the issuers are matched on the legal name. If the legal name of an issuer is 100% the same with the legal name of an organization in the internal database, then they match and a cross-reference is automatically created. This cross-reference should be uploaded to the system. The fourth step is only applicable to Moody’s. ING and Moody’s are both clients of BvD. Since BvD has Moody’s cross-references and was willing to provide them to ING, we used them for matching. If GRID’s BvD cross-reference is equal to BvD’s ID and BvD’s Moody’s cross-reference is equal to Moody’s ID, then the Moody’s issuer matches with GRID’s organization. The last step of our method is the matching through an approximate string matching algorithm. The outputs of the final step are three files. The first file contains all the issuers that are new in the internal database, the second file contains all the issuers that must be manually reconciled, because it is difficult to identify whether they are new or not, and the third file contains all the issuers that are matched with the internal organizations. For the classification of the issuers two thresholds are used: One lower (Θ1) and one upper threshold (Θ2). All the issuers, whose highest matching ratio is below the lower threshold, are stored in the first file; All the issuers, whose highest matching ratio is between the lower

and the upper threshold, are stored in the second file; All the issuers, whose highest matching ratio is higher than the upper threshold, are stored in the third file.

In our method, the emphasis is on i) improving the data quality of the internal database by matching only the correct issuers, ii) the minimization of the manual effort and iii) the minimization of the time of execution. In order to improve the speed of execution of the method, at the second step – data cleansing – we clean only three categories of the mismatches. These categories consist of the 90% of the population. The rest 10% is resolved through the approximate string matching algorithm. For data cleansing, the only available DBMS was Microsoft Access, which is not appropriate for transforming large amount of data, such as GRID. On the contrary, the approximate string matching algorithm is implemented in Python, which is much quicker than Microsoft Access. In order to decrease the wrong matches, we store the issuers that the algorithm cannot 100% recognize in the second file of the fifth step. These issuers will be manually reconciled by the user. In order to facilitate the manual reconciliation procedure and reduce the manual effort, the three highest matches are projected to the user in order to check if the issuer is new to the internal database or to select the one that it correctly matches.

5 Implementation of the method

5.1 Introduction

After finalizing the conceptual design, we implemented our method in order to examine how it works in reality and assess its results. Due to the limitations of the available DBMS, the implementation was split into two main parts. Microsoft Access 2010 was used to implement the first four steps and Portable Python 2.7.6.1 for the fifth step. The programming language in Microsoft Access was SQL using scripts, while in Portable Python the programming language was Python.

The files that were used for the implementation of the method were the same files that were used at the conceptual design. More specifically, we used the monthly files 20140504.txt, 20140501.txt and 20140701.txt from Moody's, Fitch and S&P accordingly.

5.2 Implementation of the core method in MS Access

In MS Access the analysis and manipulation of the data is done through SQL queries. There are two main types of queries: Select and Action queries. With Select queries the user can retrieve data from a table or make calculations. On the other hand, with Action queries the user can change, add or delete data. In our implementation, several Select and Action queries were used and all of them were combined through Macros. Macros were used to automate the tasks and reduce the execution time and the user involvement.

5.2.1 1st Step: Import External and Internal Data and Data Cleansing

After creating a new database, the internal and external data are imported and cleansed through two macros that contain several SQL queries. The *external data* are imported and cleansed by the macro "Issuer Import". This macro contains 4 SQL queries for Moody's and 3 SQL queries for Fitch and S&P (Table 5.1):

Query	Moody's	Fitch	S&P
a. Import Issuer	X	X	X
b. Import Translation Table	X	X	X
c. Delete MDY New	X	-	-
d. Delete Issuer Punctuation Marks	X	X	X
e. Translate Issuer	X	X	X
f. Translate Country	-	X	X

Table 5.1 - Queries per issuer of Macro "Issuer Import"

a. Import Issuer: Imports the Issuer file and stores the data at the table *Issuer.*

The query "Import Issuer" imports the issuer file in the database using the appropriate import specification. The import specification ensures that only the desired attributes of the whole file will be stored in the database. For each issuer, the attributes ID, Legal Name and Country of Residence (if applicable) are loaded and not the attributes that describe risk ratings.

b. Import Translation Table: Imports the translation table of the abbreviations and the countries.

The query "Import Translation Table" imports the translation table that contains all the abbreviations and their descriptions. In addition, the query "Import Country Table" imports the translation table for the countries in order to translate them in to ISO codes.

c. Delete MDY New: Eliminate the "(NEW)" word from the Issuer's name.

If the file is from Moody's, then the query "Delete MDY New" is executed. This query deletes all the *(NEW)* words from the legal name.

UPDATE Moodys SET Issuer_Name = REPLACE(Issuer_Name,'(NEW)','');

d. Delete Issuer Punctuation Marks: Eliminates all the punctuation marks from the legal name.

In order to remove all the punctuation marks from the legal name of the issuer, the query "Delete Issuer Punctuation Marks" is executed. This query uses the REPLACE command to replace the punctuation marks with no character.

UPDATE Moodys SET Issuer_Name = REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(Issuer_Name,'.',''),',','),'(',''),')',''),'!',''),'#',''),'$',''),'%',''),'&',''),'',''),'+',''),'/',''),':',''),';',''),'<',''),'>',''),'=',''),'?',''),'@',''),'}',''),'{',''),'_',''),'[',''),']',''),'-',''),'|',''),'\','');*

e. Translate Issuer: Translates the legal name based on the translation table.

The abbreviations are translated using the query "Translate Issuer". This query searches the legal names of the issuers for abbreviations that are stored in the translations table (Translation.Issuer) and replaces them with their description (Translation.GRID).

UPDATE Issuer, Translation SET Issuer_Name =
REPLACE(Issuer_Name,Translation.Issuer,Translation.GRID);

f. Translate Country: Translates the country based on the country table.

Last, the countries of the Issuer are translated in ISO codes using the query "Translate Country". This query replaces the countries of the issuers that are stored in the country table (Country.Description or Country. ISO3), with the ISO Codes (Country.CCRM Code).

Regarding the *internal data*, our method provides two ways for importing and cleansing. The first one is fully automated through the macro "GRID Import Automated" and the second one is semi-automated through the macro "GRID Import Semi-automated". The main difference is in the query that translates the abbreviations. The automatic macro uses the translation table to translate the abbreviations, while in the semi-automatic the user should manually insert the abbreviations and their translations into the query script. Even though the second way is not fully automated, it is much faster than the first one. This is due to the difference of the complexity of the algorithms, which is O(n) for the semi-automated way and O(m*n) for the automated one, where m is the number of records of the one file and n is the number of records of the second file. The characteristics of each way can be seen in Figure 5.3.

Query	GRID Import Automated	GRID Import Semi-automated
a. Import GRID	X	X
b. Delete GRID Punctuation Marks	X	X
c. Translate GRID Automated	X	-
d. Translate GRID Semi-Automated	-	X

Table 5.2 - Queries for GRID per Macro

a. Import GRID: Imports GRID's organizations and stores them at the table GRID.

The query "Import GRID" imports all the GRID organizations and stores them at the table GRID using the appropriate specification. The attributes of the table are: GRID Unique ID, Statutory Name and Country of Residence Code.

b. Delete GRID Punctuation Marks: Eliminates all the punctuation marks from the legal name.

In order to remove all the punctuation marks from the legal name of GRID, the query "Delete GRID Punctuation Marks" is executed. This query uses the REPLACE command to replace the punctuation marks with no character.

UPDATE GRID SET [Statutory Name] = REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE([StatutoryName],'.',''),',',''),'(',''),')',''),'!',''),'#',''),'$',''),'%',''),'&',''),'',''),'+',''),'/',''),':',''),';',''),'<',''),'>',''),'=',''),'?',''),'@',''),'}',''),'{',''),'_',''),'[',''),']',''),'-',''),'|',''),'\','');*

c. Translate GRID Automated: Translates the legal name based on the translation table.

In order to translate the legal names, the query "Translate GRID Automated" is used. The query uses the REPLACE command to replace the words that exist to the translation table. It replaces all the values of the Issuer's attribute of the Translation table with the values of the GRID's attribute.

UPDATE GRID, [Translation] SET [Statutory Name] = REPLACE([Statutory Name],Translation.Moodys,Translation.GRID);

d. Translate GRID Semi-automated: Translates the legal name without the translation table.

This method is semi-automatic as it does not load the translation table and does not translate GRID's legal names based on the values of this table. On the contrary, the user should manually edit and maintain the query that is responsible for the translation.

UPDATE GRID, [Translation] SET [Statutory Name] = REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE([Statutory Name],'Berhad','Bhd'),'Company','CO'),'Corporation','Corp'),'Government','GOVT'),'Holdings','Holding'),'Incorporated','Inc'),'International','Intl'),'Limited','Ltd'),'Public Company Limited','PCL'),'PSP','Public Sector Pension'),'Trust','TRS'),'NY','New York'),'Anonim Sirketi','AS'),'Aktiengesellschaft','AG'),'Societa Per Azioni','SPA'),'Commercial trust','COMM');

Category	Method A - Automated	Method B – Semi-automated
Advantages	Everything is done automatically.	The query for the translation has to be manually edited each time the translation table changes.
Disadvantages	Very Slow. Especially when the translation table becomes bigger.	Much quicker than method A.
Macro Name	GRID Import Automated	GRID Import Semi-automated.
Complexity	O(m*n), where m are the rows of the GRID table and n the rows of the translation table.	O(n), where n are the rows of the translation table.
Completion Time	Some hours.	1 - 2 hours.
Suggestion	It is highly recommended to use the *semi-automatic method* (Method B), as it runs much faster than method A. The query is not complicated and no programming skills are required to maintain it.	

Table 5.3 - Comparison of Automated and Semi-automated methods

5.2.2 2nd Step: Match on Cross-Reference

The matching procedure is executed through the macro "Recon". Recon executes eleven queries for Fitch and S&P and seventeen for Moody's that create new tables, update the contents of existing tables and export data in excel files. For Moody's, Recon contains additional queries that are related to the matching procedure using BvD's cross-references. The queries for each issuer are shown in Table 5.1. For Fitch and S&P, the unmatched issuers for the whole matching procedure are stored in the file Unmatched_LN, while for Moody's they are stored in the file Unmatched_BvD.

Query	Moody's	Fitch	S&P
001 Create Matched_ID table	X	X	X
002 Update Matched_ID	X	X	X
003 Create Unmatched_ID table	X	X	X
004 Update Unmatched_ID	X	X	X
Export Matched_ID	X	X	X
005 Create Matched_LN table	X	X	X
006 Update Matched_LN	X	X	X
007 Create Unmatched_LN table	X	X	X
008 Update Unmatched_LN	X	X	X
Export Matched_LN	X	X	X
Export Unmatched_LN		X	X
Import BvD-Moody's	X	-	-
Import GRID-BvD	X	-	-
009 Create Unmatched_BvD table	X	-	-
010 Update Unmatched_BvD	X	-	-
Export Matched_BvD	X	-	-
Export Unmatched_BvD	X	-	-

Table 5.4 - Queries of Recon macro per Issuer

In this step, the issuers are matched to the GRID's organization based on GRID's cross-reference. If the ID of an issuer is equal to the cross-reference of a GRID organization, then they match and the issuer is stored in the table "Matched_ID". If it doesn't match, the issuer is stored in the table "Unmatched_ID". For the identification of the matched issuers, the query *002 Update Matched_ID* is used and for the identification of the unmatched issuers the query *004 Update Unmatched_ID*.

Matched_ID:

INSERT INTO Matched_ID (GRID_ID, GRID_Name, Issuer_Name, Issuer_ID, GRID_Country, Issuer_Country) SELECT GRID_XREF.[GRID Unique ID], GRID_XREF.[Statutory Name], Issuer.Issuer_Name, GRID_XREF.[Cross-reference], GRID_XREF.[Country of Residence Code], Issuer.Country FROM Issuer INNER JOIN GRID_XREF ON GRID_XREF.[Cross-reference] = Issuer.Issuer_ID;

Unmatched_ID:

INSERT INTO Unmatched_ID (Issuer_ID, Issuer_Name, Issuer_Country) SELECT Issuer_ID, Issuer_Name FROM Issuer WHERE Issuer.Issuer_ID NOT IN (SELECT Matched_ID.Issuer_ID FROM Matched_ID);

The Matched_ID table contains the following attributes: GRID Unique Identifier, GRID Statutory Legal Name, Issuer Statutory Legal Name, Issuer Unique Identifier, GRID Country of Residence and Issuer Country of Residence. The Unmatched_ID table contains the same attributes with the Issuer file.

5.2.3 3rd Step: Match on Legal Name

In this step, the remaining unmatched issuers (Unmatched_ID) are matched on the statutory legal name. If the legal name of an issuer is exactly the same as the legal name of a GRID organization, then they match and the issuer is stored in the table "Matched_LN". If it doesn't match, the issuer is stored in the table "Unmatched_LN". For the identification of the matched issuers, the query *006 Update Matched_LN* is used and for the identification of the unmatched issuers the query *008 Update Unmatched_LN*. When an issuer is matched, then a cross-reference for GRID is automatically created and stored. The value of the cross-reference is the same with the issuer's ID.

Matched_LN:

INSERT INTO Matched_LN (GRID_ID, GRID_Name, Issuer_NAME, XREF, GRID_Country, Issuer_Country) SELECT GRID.[GRID Unique ID], GRID.[Statutory Name], Unmatched_ID.Issuer_Name, Unmatched_ID.MDY_ID, GRID.[Country of Residence Code], Unmatched_ID.Issuer_Country FROM Unmatched_ID, GRID WHERE Unmatched_ID.Issuer_Name = GRID.[Statutory Name];

Unmatched_LN:

INSERT INTO Unmatched_LN (Issuer_ID, Issuer_Name, Issuer_Country) SELECT Issuer_ID, Issuer_Name, Issuer_Country FROM Issuer WHERE Issuer.Issuer_ID NOT IN (SELECT Matched_ID.Issuer_ID FROM Matched_ID);

The Matched_LN table contains the following attributes: GRID Unique Identifier, GRID Statutory Legal Name, Issuer Statutory Legal Name, GRID's Issuer Cross-reference, GRID Country of Residence and Issuer Country of Residence. The Unmatched_LN table contains the same attributes with the Issuer file.

5.2.4 4th Step: Match using BvD cross-references (only for Moody's)

First of all, the file of BvD's Moody's cross-references is imported to the database and is stored in BvD9 table. This table contains only two attributes: BvD9 number and Moody's Issuer Number. In addition, the file of Moody's BVD cross-references is imported to the database and is stored in GRID BVD table. This table also contains two attributes: GRID Unique Identifier and BvD Cross-reference. Afterwards, the remaining unmatched issuers (Unmatched_LN) are matched using BvD's cross-references. If ING's BvD cross-reference is equal to BvD ID and BvD's Moody's cross-reference is equal to Moody's ID, then they match and the issuer is stored in the table "Matched_BvD". If it does not match, the issuer is stored in the table "Unmatched_BvD". For the identification of the matched issuers, the query *008 Update Matched_BvD* is used and for the identification of the unmatched issuers the query *010 Update Unmatched_BvD*. When an issuer is matched, then a cross-reference for GRID is automatically created and stored. The value of the cross-reference is the same with the issuer's ID.

Matched_BvD:

INSERT INTO Matched_BvD (GRID_ID, MDY_ID, MDY_Issuer_Name) SELECT [GRID BVD].[GRID Unique ID], BVD9.[BvD9 number], Unmatched_LN.Issuer_Name FROM [GRID BVD], BVD9, Unmatched_LN WHERE (BVD9.[BvD9 number]=[GRID BVD].[Cross-reference] AND BVD9.[Moody's Issuer Number]=Unmatched_LN.Issuer_ID);

Unmatched_BvD:

INSERT INTO Unmatched_BvD (MDY_ID, MDY_Issuer_Name) SELECT Issuer_ID, Issuer_Name FROM Unmatched_LN WHERE Unmatched_LN.Issuer_ID NOT IN (SELECT Matched_BvD.MDY_ID FROM Matched_BvD);

The Matched_BvD table contains the following attributes: GRID Unique Identifier, GRID's Issuer Cross-reference and Issuer Statutory Legal Name. The Unmatched_BvD table contains the same attributes with the Issuer file.

After implementing the four first steps of our method, we applied them to the monthly files 20140504.txt, 20140501.txt and 20140701.txt from Moody's, Fitch and S&P accordingly. The results, that are shown in Table 5.5, show that the majority of the records for every issuer were matched on the Cross-reference and less than 10% of the issuers did not match at any step.

	Moody's	Fitch	S&P
Total Records	10971	7411	10104
Matched_ID	8945	7044	9070
Unmatched_ID	2026	367	1034
Matched_LN	963	85	413
Unmatched_LN	1063	282	621
Matched_BvD	203	-	-
Unmatched_BvD	860	-	-

Table 5.5 - Results of matches after applying the first four steps

5.3 Implementation of the approximate string matching algorithms in Python

Since a significant amount of issuers (on average 6%) were not matched in any of the first four steps, an approximate string matching algorithm is applied at the last step of our method. Due to the fact that Visual Basic for Applications (VBA) in Microsoft Access 2010 is slow and cannot deal with large amount of data, such as comparing the remaining unmatched issuers to GRID's organization, our method was implemented in Python language using Python Portable 2.7.6.1. The main reason why Python Portable was chosen instead of the full desktop Python environment was because we did not have the administration rights to install the desktop versions at the PC in ING. On the contrary, the Portable version did not require any administration rights and its functionality is exactly the same with the desktop

version. More specifically, the python script was programmed in PyCharm-Portable.exe[10]. This is a user-friendly Python Integrated Development Environment (IDE) that provides unique code assistance, such as finding and installing packages without using the command prompt.

In the literature, as mentioned in Chapter 3.4, there are many approximate string matching algorithms. Two of the most common and widely used ones are Levenshtein Distance Algorithm [Levenshtein 1966] and Jaro-Distance Algorithm. The Levenshtein Distance Algorithm according to the literature is more suitable for small and large strings, while the Jaro-Distance Algorithm is more suitable for small and medium strings. In order to identify which one produces better results, both were implemented and evaluated. In the beginning both were implemented from scratch in python language and it worked perfectly for a small sample data set. When it was applied for the whole GRID file, the program crashed as it run out of memory. The main reason behind this occurrence was that Python does not release the memory of the system that was allocated for another reason. The solution of this problem was the use of Python C extension modules that could allocate and release the memory. As a result, the package "python-Levenshtein version 0.11.2"[11] was installed using the Project Interpreter. In this package, both the Levenshtein Distance and Jaro-Distance were included through the corresponding function. For example, the ratio of the Levenshtein Distance and the Jaro-Distance between two strings, A and B, is calculated by the following commands:

$$Levenshtein\ Ratio = Levenshtein.ratio(A, B) * 100 \quad \text{Eq. 5}$$

$$Jaro\ Ratio = Levenshtein.jaro(A, B) * 100 \quad \text{Eq. 6}$$

As it was mentioned in Chapter 4.3, there will be two thresholds, one lower and one upper, based on which the issuers will be categorized to Unmatched (New), Unmatched (Manual) and Matched. For the computation of the appropriate value of the thresholds the monthly S&P file was selected, because Moody's does not provide the country of residence of the issuer and because the S&P's file is bigger and contains higher variety of mismatches. For the definition of the thresholds, all the remaining unmatched issuers from S&P were manually checked whether they exist in the internal database and afterwards different thresholds were applied to estimate which one produces the best results. The focus was, first of all, on maximizing the precision of the algorithm and secondly on minimizing the manual effort for reconciliation. At the beginning, only the statutory legal name was included in the analysis using the Levenshtein Distance algorithm. Based on the results in Table A.1 in Appendix B, the optimal lower threshold is 80 and the optimal upper threshold is 97. All the issuers whose highest match is lower than 80 will be stored in the file Unmatched (New) file, all the issuers whose highest match is higher than 79 and lower than 97 will be stored in the Unmatched (Manual) file and all the issuers whose highest match is higher than 96 will be stored in the Matched file. With these thresholds the wrong matched are only 0.16 % and the manual effort is 71.66%. Since the issuers' statutory legal name is the same for all the risk raters, these thresholds will be applied to both Fitch and Moody's. In addition, we

[10] http://www.jetbrains.com/pycharm/
[11] https://pypi.python.org/pypi/python-Levenshtein/0.11.2

observed that many issuers from Moody's and Fitch that exist in the internal database had a highest match of 80 and almost none less than 80. Thus, if we had changed the values of the thresholds, then the precision of the algorithm would have dropped.

The next step of our analysis was to include the country of residence, when applicable, to the matching process. Based on our observations and after experimenting with the numbers, we concluded that if the country of residence between an issuer and an organization in the internal database is not the same, if we reduce the matching ratio by 10% (eq. 7), then many of the issuers that fell into the Unmatched (Manual) category and do not exist in GRID, will fall into the Unmatched (New) category without affecting the precision of the algorithm. As a result, the manual effort will significantly be decreased by 12.56% (Table 5.6).

If CountryOfResidence(Issuer) != CountryOfResidence(GRID) then{

*Ratio = 0.9 * Ratio* *Eq. 7*

} End IF

	Without Country	**With Country**
Total	621	621
Correct	175	251
Correct %	28,18%	40,42%
False	1	3
False %	0,16%	0,48%
Manual Effort	445	367
Manual Effort %	71,66%	59,10%

Table 5.6 - Comparison of S&P's results of the thresholds 80 and 97 with and without the country of residence

After defining the optimal thresholds based on our objectives and the appropriate formula for the countries of residence, the same logic was applied to the monthly files of Moody's and Fitch. Due to the fact that Moody's did not provide the country of residence, the Equation 7 was not included in the python script. In addition, the Jaro-Distance algorithm was applied for Moody's, Fitch and S&P in order to compare the results of the two algorithms and select the most suitable one. The results of the approximate string matching algorithms can be seen in Table 5.7.

The first six rows indicate the approximate string algorithm, the risk rater, whether the country of residence is included in the matching process or not, the total number of records of the issuer files, the low threshold and the upper threshold. The category 'New Organizations' contains four subcategories. The 'Records' subcategory shows the number of issuers that exist in this category and are labeled as new, the 'Correct' subcategory shows the number of issuers that were correctly identified as new, the 'False (Exist)' subcategory shows the number of issuers that have been matched with the correct organization but exist in GRID and the 'False (Wrong match)' shows the number of issuers that have been matched with a wrong organization and exist in GRID.

The category ‘Manual Reconciliation’ contains also four subcategories. The ‘Records’ subcategory shows the number of issuers that exist in this category and should be manually reconciled, the ‘Correct’ subcategory shows the number of issuers that have been matched with the correct organization, the ‘False (New)’ subcategory shows the number of issuers that do not exist in GRID and the ‘False (Wrong match)’ subcategory shows the number of issuers that have been matched with a wrong organization.

The category ‘Matched’ contains also four subcategories. The ‘Records’ subcategory shows the number of issuers that exist in this category and have been labeled as correctly matched, the ‘Correct’ subcategory shows the number of issuers that have been matched with the correct organization, the ‘False (New)’ subcategory shows the number of issuers that do not exist in GRID but have been matched with an organization and the ‘False (Wrong match)’ subcategory shows the number of issuers that have been matched with a wrong organization.

Last but not least, the category ‘Total’ contains also four subcategories and gives a summary of the previous categories. The ‘Correct’ subcategory shows the number of issuers that have been correctly assigned to the categories, the ‘False’ subcategory shows the number of issuers that have been incorrectly assigned to the categories and the ‘Manual Effort’ subcategory shows the number of issuers that should be manually reconciled

Algorithm		**Levenshtein**				**Jaro Distance**		
Risk Rater		**Fitch**	**SAP**		**Moody's**	**Fitch**	**SAP**	**Moody's**
Country		*Included*	*Included*	*Excluded*	*Excluded*	*Included*	*Included*	*Excluded*
Records		282	621	621	860	282	621	860
Low Threshold		80	80	80	80	80	80	80
Upper Threshold		97	97	97	97	97	97	97
New Organizations	Records	141	238	171	211	110	149	62
	Correct	138	235	170	206	108	108	59
	Correct %	97,87%	98,74%	99,42%	97,63%	98,18%	72,48%	95,16%
	False (Exist)	3	1	1	0	1	3	0
	False (Exist) %	2,13%	0,42%	0,58%	0,00%	0,91%	2,01%	0,00%
	False (Wrong match)	0	2	0	5	1	1	3
	False (Wrong match) %	0,00%	0,84%	0,00%	2,37%	0,91%	0,67%	4,84%
	False	3	3	1	5	2	4	3
	False %	2,13%	1,26%	0,58%	2,37%	1,82%	2,68%	4,84%
Manual Reconciliation	Records	134	367	445	539	166	456	687
	Correct	24	88	95	56	25	77	53
	Correct %	17,91%	23,98%	21,35%	10,39%	15,06%	16,89%	7,71%
	False (New)	107	277	341	475	134	366	616
	False (New) %	79,85%	75,48%	76,63%	88,13%	80,72%	80,26%	89,67%
	False (Wrong match)	3	2	9	8	7	13	18
	False (Wrong match) %	2,24%	0,54%	2,02%	1,48%	4,22%	2,85%	2,62%
Matched	Records	7	16	5	110	6	14	111
	Correct	7	16	5	94	3	12	92
	Correct %	100,00%	100,00%	100,00%	85,45%	50,00%	85,71%	82,88%
	False (New)	0	0	0	15	3	2	18
	False (Exist) %	0,00%	0,00%	0,00%	13,64%	50,00%	14,29%	16,22%
	False (Wrong match)	0	0	0	1	0	0	1
	False (Wrong match) %	0,00%	0,00%	0,00%	0,91%	0,00%	0,00%	0,90%
	False	0	0	0	16	3	2	19
	False %	0,00%	0,00%	0,00%	14,55%	50,00%	14,29%	17,12%
Total	Total	282	621	621	860	282	619	860
	Correct	145	251	175	316	114	122	170
	Correct %	51,42%	40,42%	28,18%	36,74%	40,43%	19,65%	19,77%
	False	3	3	1	21	5	6	22
	False %	1,06%	0,48%	0,16%	2,44%	1,77%	0,97%	2,56%
	Manual Effort	134	367	445	539	166	456	687
	Manual Effort %	47,52%	59,10%	71,66%	62,67%	58,87%	73,43%	79,88%

Table 5.7 - Results of the approximate string matching algorithms for all files

Based on the results, it is observed that when the country of residence is included in the string matching process the manual effort is significantly reduced and the results are overall better. For example, the results of the matching procedure of the Moody's file are much lower to the results of Fitch and S&P due to the absence of the country of residence. Moreover, for this kind of data Levenshtein Distance algorithm produces more accurate and precise predictions than Jaro-Distance algorithm and therefore is preferred. As it is shown in Table 5.8, the average length of the statutory legal names of the issuers is 27, which is relatively long. Thus, it was expected that Levenshtein Distance will have better results than Jaro-Distance, because Jaro-Distance performs better on short strings.

Legal Name	Moody's	Fitch	S&P	Total
Max Length	50	114	81	114
Min Length	5	5	4	4
Average Length	25	31	26	27

Table 5.8 - Length of the legal names of the unmatched issuers

For the measurement and the comparison of the accuracy of the two algorithms, the precision, recall and F1-score performance indicators were calculated. Since these performance indicators were also used for the validation of our method, more details about them are explained in Chapter 6.1. Based on the results in Table 5.9, it is observable that the Levenshtein Distance algorithm is more accurate and precise than Jaro-Distance algorithm.

	Levenshtein Distance Algorithm			Jaro-Distance Algorithm		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
S&P	98,11%	97,20%	97,65%	87,50%	95,70%	91,42%
Fitch	91,18%	91,18%	91,18%	81,58%	93,33%	87,06%
Moody's	94,86%	99,34%	97,05%	89,62%	97,97%	93,61%

Table 5.9 - Precision, Recall and F1 of the algorithms per issuer

In theory, Jaro-Distance algorithm performs better that Levenshtein Distance algorithm for small strings. Thus, a combination of the two algorithms, Jaro-Distance for short strings and Levenshtein Distance for long strings, could have provided better results. Therefore, we monitored the performance of Levenshtein Distance algorithm for small strings. In our case, a small string consists of less than 9-11 characters. The results of our analysis that are shown in table 5.10 and 5.12 show that the Levenshtein Distance algorithm for the legal names of S&P and Moody's has 100% precision, recall and F1-score. For Fitch (Table 5.11), there were no issuers that exist in GRID and the length of its legal name is less than 12 characters. Therefore, the approximate string matching algorithm that our method will use is only the Levenshtein Distance algorithm.

String Length	Records	New	Correct	False	Precision	Recall	F1
<12	27	21	6	0	100%	100%	100%
<11	22	18	4	0	100%	100%	100%
<10	12	11	1	0	100%	100%	100%

Table 5.10 - Performance of Levenshtein Distance algorithm on short legal names for S&P

String Length	Records	New	Correct	False	Precision	Recall	F1
<12	12	12	0	0	-	-	-
<11	8	8	0	0	-	-	-
<10	5	5	0	0	-	-	-

Table 5.11 - Performance of Levenshtein Distance algorithm on short legal names for Fitch

String Length	Records	New	Correct	False	Precision	Recall	F1
<12	33	28	5	0	100%	100%	100%
<11	22	18	4	0	100%	100%	100%
<10	15	12	3	0	100%	100%	100%

Table 5.12 - Performance of Levenshtein Distance algorithm on short legal names for Moody's

Functionality Python Script:

The Python script in Appendix C is the final version that includes the country of residence (lines 44-45) in the matching process and uses the Levenshtein Distance algorithm to calculate the ratio (lines 40-42). The lines 44-45 are only used when the risk rater provide the country of residence of the issuer. For every issuer, the three highest matches are calculated and sorted in an ascending order. The variables where the ratios are stored are maxRatio1, maxRatio2 and maxRatio3. The maxRatio3 is bigger or equal to the maxRatio2 and the maxRatio2 is bigger or equal to the maxRatio1. For each organization that has these ratios, we also store the ID, Legal Name and Country of Residence. In order to sort the ratios and store only the three highest matches for each issuer, the following algorithm is used:

```
Calculate NewRatio;
If (NewRatio > maxRatio3) then{
        maxRatio1 = maxRatio2;
        maxRatio2 = maxRatio3;
        maxRatio3 = NewRatio;
}else if (NewRatio > maxRatio2) then{
        maxRatio1 = maxRatio2;
        maxRatio2 = newRatio;
}else if (NewRatio > maxRatio1) then{
        maxRatio1 = NewRatio;
}end if
```

If the highest match (maxRatio3) of an issuer is below 80, then only the internal organization with the highest match is stored in the file Unmatched (New). If the highest match is between 80 and 96, then the three highest matches are stored in the file Unmatched (Manual) and if the highest match is above 96, then only the internal organization with the highest match is stored in the file Matched. These output files have a .txt format and for each record the following attributes are stored: Issuer ID, Issuer Legal Name, GRID Legal Name, GRID ID, Issuer Country of Residence, GRID Country of Residence and the Percentage

of Match. All the issuers that are stored in the file Unmatched (New) do not exist in the internal database and the user should not investigate them for the reconciliation. On the other hand, if the owner of the database could upload these issuers to the database in order to enrich it. However, this is not recommended, because the data from Moody's, Fitch and S&P are not clean and because additional information should be included when creating a new issuer, such as customer type, country of incorporation, town of residence etc. All the issuers that are stored in the file Unmatched (Manual) should be manually reconciled by the user. In order to facilitate the manual reconciliation process and decrease the manual effort, we project the three highest matches for every issuer and the percentage, legal name and country of residence of each matching pair. In that way, if the organization exists in GRID, it is very likely that it will be one of these three organizations. Thus, the user will be able to quickly identify the correct one. Last but not least, all the issuers that are stored in the file Matched are correctly matched and automatically a cross-reference is created. The user should manually check them, but he should upload the cross-references into the database. In that way, the following reconciliation process will be quicker as more issuers will be matched on the second step of the method (Match on Cross-reference).

SAP_ID	SAP_LName	GRID_LName	GRID_ID	SAP_Country
308837	People's United Financial Inc	Peoples United Financial Inc	36350803	US
403441	Credit und Volksbank eG	Creditund Volksbank eG	36140321	DE
403980	Volksbank Suedheide eG	Volksbank Sudheide eG	36397999	DE
422344	Financiere Gaillon 8 SA	Financiere Gaillon 8 SAS	43625082	FR
454981	Servicios Corporativos Javer SAPI de C V	Servicios Corporativos Javer SA PI de CV	43500139	MX
495581	Globoaves Sao Paulo Agroavícola Ltda	Globoaves Sao Paulo Agroavicola Ltda	47722444	BR
522560	Vietnam Export Import Commercial Joint Stock Bank	Vietnam ExportImport Commercial JointStock Bank	36002924	VN
538819	Corporacion Azucarera del Peru S A	Corporacion Azucarera del Peru SA	46894682	PE
543368	Infraestructura Energetica Nova SA de CV	Infraestructura Energetica Nova SAB de CV	47204483	MX

Picture 5.1- Example of Matched File

The process in total when using the semi-automatic way takes approximately three and a half hours, but when using the automatic way it takes more than twelve hours. Thus, it is recommended to use the semi-automatic way as it is much faster than the automatic one.

5.4 Conclusion

The implementation of our method consists of two main parts. The first part was implemented in Microsoft Access 2010 in SQL scripts. At the first step, the internal and external files were imported into Access' database, and the statutory legal names and the countries of residence were cleaned. The cleaning procedure consists of the removal of the punctuation marks inside the legal names, the translation of the abbreviations of the legal names and the translation of the countries of residence to ISO codes. At the second step, the issuers were matched on the existing cross-references and the remaining unmatched issuers were matched at the third step on the legal name. For Moody's our method has an additional step, which uses a third party, BvD, to match the issuers. BvD has cross-references for Moody's and in GRID there are cross-references for BvD's organizations. As a result, the remaining unmatched issuers are matched through BvD. The second part was implemented in Portable Python 2.7.6.1 using Python scripts. The remaining unmatched issuers were matched using Levenshtein Distance algorithm. The outputs of this part are three files: Unmatched (New), Unmatched (Manual) and Matched. The first file contains all the issuers

that were identified as new, the second file contains all the issuers that have to be manually reconciled by the user and the third file contains all the issuers that are correctly matched.

The final results of our method for the monthly files 20140504.txt, 20140501.txt and 20140701.txt from Moody's, Fitch and S&P accordingly are shown in Table 5.13 and 5.14. On average, 94.49% issuers were matched and the required manual effort was only 3.45%. Taken into account that the user can select between only three organizations in order to identify with which one the issuer matches the manual effort and time drops even more.

	Moody's	Fitch	S&P
Total Records	10.971	7.411	10.104
Matched_ID	8.945	7.044	9.070
Unmatched_ID	2.026	367	1.034
Matched_LN	963	85	413
Unmatched_LN	1.063	282	621
Matched_BvD	203	-	-
Unmatched_BvD	860	-	-
Unmatched (New)	211	141	238
Unmatched (Manual)	539	134	367
Matched	110	7	16

Table 5.13 - Results after implementing the method to Moody's, Fitch and S&P monthly files

	Moody's		Fitch		S&P		Average
Total Records	10.971	100%	7.411	100%	10.104	100%	100%
Matched	10.221	93.16%	7.136	96.29%	9.499	94.01%	94.49%
Manual Effort	539	4.91%	134	1.81%	367	3.63%	3.45%
New Organizations	211	1.92%	141	1.9%	238	2.36%	2.06%

Table 5.14 - Number of matched and unmatched issuers after implementing the method to Moody's, Fitch and S&P monthly files

6 Method validation

6.1 Validation method

After designing and implementing our method in Chapter 4 and 5, we used one monthly file for each risk rater in order to verify the results of our method. Even though these files could have been used to validate our method, another data set of three different monthly files was used to prove our method's validity. The reasoning behind this decision was the exclusion of the overfitting phenomenon. The most recent files were selected, because they would contain new issuers that did not exist in the previous ones. The monthly files were preferred to the daily files, because the contained more records and thus had greater variety.

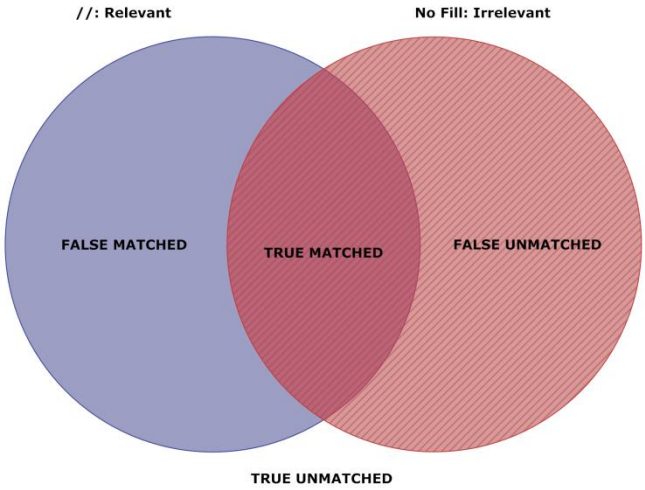
Two types of validation indicators were used. For the first type of validation, the precision, recall and F1 score were calculated for each file and for each approximate string matching algorithm. These performance indicators are widely used in evaluating search strategies [Goutte et al. 2005; Levin et al. 2012] in order to measure the search effectiveness. The *precision* measures the ratio of the number of true matched issuers to the total number of false matched and true matched issuers (eq. 8). The *recall* measures the ratio of the number of true matched issuers to the total number of false unmatched and true matched issuers (eq. 9). The *F1 score* measures the test's accuracy and is the harmonic mean of the precision and recall (eq. 10). Prior to the calculation of the true matched, false matched, true unmatched and false unmatched issuers, we excluded all the issuers that did not exist in GRID.

$$Precision = \frac{True\ Matched}{False\ Matched + True\ Matched} \quad \text{Eq. 8}$$

$$Recall = \frac{True\ Matched}{False\ Unmatched + True\ Matched} \quad \text{Eq. 9}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \quad \text{Eq. 10}$$

The second type of validation indicator was the total number of matched issuers, the percentage of the manual effort and the total number of the issuers that do not exist in GRID. In order to calculate the total number of matched issuers, we added the matched issuers from the files Matched_ID, Matched_LN, Matched_BvD and Matched. For the calculation of the percentage of the manual effort, we computed the ratio of the number of the issuers that were included in the Unmatched (Manual) file to the total number of the issuers of the original file. Last but not least, for the calculation of the total number of the issuers that did not exist in GRID, we counted the records of the Unmatched (New) file. For both types of validation indicators, we compared the results of the newest data set to the older data set that was used in Chapter 4 and 5. If the numbers are at the same level, then the validity of our method will be proven.



Picture 6.1 - Relevant and irrelevant matched and unmatched issuers

6.2 Results of validation

The files that were randomly selected and used for the validation of the method were the monthly files 20140611.txt, 20140801.txt and 201407801.txt from Moody's, Fitch and S&P accordingly. For every issuer, we checked manually using GRID's search engine to see if it exists in GRID or not.

		Levenshtein			Jaro Distance		
		Fitch	**SAP**	**Moody's**	**Fitch**	**SAP**	**Moody's**
Country		*With*	*With*	*Without*	*With*	*With*	*Without*
Records		345	698	854	345	698	854
Low Threshold		80	80	80	80	80	80
Upper Threshold		97	97	97	97	97	97
	Records	159	267	225	133	170	61
	Correct	155	264	221	131	165	60
	Correct %	97,48%	98,88%	98,22%	98,50%	97,06%	98,36%
	False (Exist)	4	1	0	1	5	0
New Organizations	False (Exist) %	2,52%	0,37%	0,00%	0,75%	2,94%	0,00%
	False (Wrong match)	0	2	4	1	1	1
	False (Wrong match) %	0,00%	0,75%	1,78%	0,75%	0,59%	1,64%
	False	4	3	4	2	6	1
	False %	2,52%	1,12%	1,78%	1,50%	3,53%	1,64%
	Records	177	415	582	205	514	744
	Correct	30	88	58	33	81	56
	Correct %	16,95%	21,20%	9,97%	16,10%	15,76%	7,53%
	False (New)	144	325	515	165	423	671
Manual Reconciliation	False (New) %	81,36%	78,31%	88,49%	80,49%	82,30%	90,19%
	False (Wrong match)	3	2	9	7	10	17
	False (Wrong match) %	1,69%	0,48%	1,55%	3,41%	1,95%	2,28%
	Manual Effort	177	415	582	205	514	744
	Records	9	16	47	7	14	49
	Correct	8	16	32	3	12	26
	Correct %	88,89%	100,00%	68,09%	42,86%	85,71%	53,06%
	False(New)	1	0	14	4	2	2
Matched	F False (Exist) %	11,11%	0,00%	29,79%	57,14%	14,29%	4,08%
	False (Wrong match)	0	0	1	0	0	1
	False (Wrong match) %	0,00%	0,00%	2,13%	0,00%	0,00%	2,04%
	False	1	0	15	4	2	21
	False %	11,11%	0,00%	31,91%	57,14%	14,29%	6,12%
	Total	**345**	**698**	**854**	**345**	**698**	**854**
	Correct	**164**	**280**	**268**	**138**	**179**	**109**
	Correct %	**47,54%**	**40,11%**	**31,38%**	**40,00%**	**25,64%**	**12,76%**
Total	**False**	**5**	**3**	**19**	**6**	**8**	**22**
	False %	**1,45%**	**0,43%**	**2,22%**	**1,74%**	**1,15%**	**2,58%**
	Manual Effort	**177**	**415**	**582**	**205**	**514**	**744**
	Manual Effort %	**51,30%**	**59,46%**	**68,15%**	**59,42%**	**73,64%**	**87,12%**

Table 6.1 - Results of Levenshtein Distance Algorithm for all files

	Fitch		S&P		Moody's		Fitch		S&P		Moody's	
	Old File	*New File*	*Old File*	*New File*	*Old File*	*New File*	*Old File*	*New File*	*Old File*	*New File*	*Old File*	*New File*
Total	282	345	621	698	860	854	282	345	619	698	860	854
Correct	145	164	251	280	316	268	114	138	122	179	170	109
Correct%	51,4%	47,5%	40,4%	40,1%	36,7%	31,3%	40,4%	40,0%	19,6%	25,6%	19,7%	12,7%
FALSE	3	5	3	3	21	19	5	6	6	8	22	22
False %	1,06%	1,45%	0,48%	0,43%	2,44%	2,22%	1,77%	1,74%	0,97%	1,15%	2,56%	2,58%
Manual Effort	134	177	367	415	539	582	166	205	456	514	687	744
Manual Effort %	47,5%	51,3%	59,1%	59,4%	62,6%	68,1%	58,8%	59,4%	73,4%	73,6%	79,8%	87,1%

Table 6.2 - Comparison of the results of the old and new/validation files per risk rater per algorithm

In Table 6.1, it is observed that the predictions using the Levenshtein Distance algorithm are indeed more accurate and precise than the predictions of the Jaro-Distance algorithm. The false predictions with the Jaro-Distance algorithm are increased by 33.3% and the required manual effort is increased by 24.6%, while the correct predictions are decreased by 40.2%. In addition, the manual effort for the new files is almost the same with the false records for the old files. On the other hand, the correct records and the manual effort have slightly increased for the new files, but they both remain on the same levels. Consequently, this small difference is due to the difference of the records of each file and not due to overfitting.

In order to verify the results of our method, we calculated the precision, recall and F1 score performance indicators using the Levenshtein Distance Algorithm (Table 6.2). Our method using Levenshtein Distance algorithm gives higher values of precision, recall and F1 score than using the Jaro-Distance, even though Jaro-Distance have also high precision, recall and F1. In Table 6.4, it is observed that the precision, recall and F1 score of the different files have similar high values, which proves the validity of our method.

	Levenshtein Distance Algorithm			**Jaro-Distance Algorithm**		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*
S&P	98,11%	97,20%	97,65%	90,29%	93,94%	92,08%
Fitch	92,68%	90,48%	91,57%	83,72%	94,74%	88,89%
Moody's	90,00%	95,74%	92,78%	82,00%	98,80%	89,62%

Table 6.3 - Precision, Recall and F1 of the algorithms per issuer

Levenshtein Distance Algorithm						
	Precision		*Recall*		*F1*	
	New	Old	New	Old	New	Old
S&P	98,11%	98,11%	97,20%	95,41%	97,65%	97,65%
Fitch	92,68%	91.18%	91,18%	83,78%	91,57%	91,18%
Moody's	90,00%	94.86%	99,34%	91,46%	92,78%	97,05%
Jaro-Distance Algorithm						
	Precision		*Recall*		*F1*	
	New	Old	New	Old	New	Old
S&P	90,29%	87,50%	93,94%	95,70%	92,08%	91,42%
Fitch	83,72%	81,58%	94,74%	93,33%	88,89%	87,06%
Moody's	82,00%	89,62%	98,80%	97,97%	89,62%	93,61%

Table 6.4 - Comparison of the Precision, Recall and F1 of the old and the new files per algorithm per issuer

6.3 Conclusion

For the validation of our method, different monthly filed were used for each risk rater. After manually checking whether the issuers of these files exist in the internal database or not, the precision, recall and F1 score were calculated. The precision, recall and F1 show that the Levenshtein Distance algorithm is indeed better that the Jaro-Distance algorithm, even though Jaro-Distance has also high precision, recall and F1 score values. Furthermore, the results of the new files are similarly high compared with the results of the files that were used in Chapter 5. This proves the validity of our method and that the phenomenon of overfitting was largely avoided.

The final results of our method for the monthly files 20140611.txt, 20140801.txt and 20140801.txt from Moody's, Fitch and S&P accordingly are shown in Table 5.8 and 5.9. On average, 93.79% issuers were matched and the required manual effort was only 3.92%. For the reconciliation of the 3.92% of the file, our method gives the user the three highest matches for each issuer. The user can select the correct match only from these three organizations and not from the entire database. Thus, the manual effort is further reduced.

	Moody's	Fitch	S&P
Total Records	10.971	7.410	10.180
Matched_ID	8.945	6.931	9.035
Unmatched_ID	2.026	479	1.145
Matched_LN	1.005	134	447
Unmatched_LN	1.021	345	698
Matched_BvD	167	-	-
Unmatched_BvD	854	-	-
Unmatched (New)	225	159	267
Unmatched (Manual)	582	177	415
Matched	47	8	16

Table 6.5 - Results after implementing the method to different Moody's, Fitch and S&P monthly files

On average, using the new files 93.79% of the issuers were matched in comparison to the 94.49% of the old files. The manual effort was slightly increased from 3.45% to 3.92% of the issuers. In general, the number of matched issuers and the manual effort remained at the same levels for both the new and old files.

	Moody's		Fitch		S&P		Average
Total Records	10.971	100%	7.411	100%	10.180	100%	100%
Matched	10.164	92.64%	7.073	95.44%	9.498	93.3%	93.79%
Manual Effort	582	5.3%	177	2.39%	415	4.08%	3.92%
New Organizations	225	2.05%	159	2.15%	267	2.62%	2.27%

Table 6.6 - Number of matched and unmatched issuers after implementing the method to different Moody's, Fitch and S&P monthly files

7 Conclusion

7.1 Discussion

An automated method to reconcile risk rating data from credit risk agencies is developed in this thesis. The input of the method was based on ING's internal data and on the issuer file data that Moody's, Fitch and Standard and Poor's are providing to ING. The method was implemented using ING's infrastructure. Even though our method was based on ING's data, it can be applied to other organizations, because the risk rating data are the same in any organization. More specifically, we only used the statutory legal name and the country of residence of each issuer in the reconciliation process, which are both the same in any organization. In addition, the Big Three agencies control approximately 95% of the rating market share worldwide, which means that the external data in the reconciliation process will be 100% the same for every organization. Prior to the design, we analyzed and categorized the reasons of mismatches between the internal and external issuers. The majority of the categories were also detected during the literature review, which indicates that our method can be generalized.

In answering to our sub-research questions, the main challenge was to clean the external and internal data in a way that they could be easily matched. For that reason, we manually analyzed the reasons of mismatches for an entire issuer file per agency using ING's GRID search engine. The analysis showed that the majority of the mismatches were due to the fact that:

i. The issuers did not exist in ING's GRID database;
ii. The issuers had different punctuation marks with the organizations in GRID;
iii. The issuers or the organizations in GRID had abbreviated words.

On average, these categories consist of 89% of the total population, which is a relatively large percentage. In Fitch's and S&P's file, the majority of the issuers did not exist in GRID, while in the Moody's file half of the issuers did not exist in GRID. During the analysis of the data, all the abbreviations that were met were stored in a translation table. These abbreviations were later used in our methodology.

Our method was designed based on the results of the data analysis and in a way that could be implemented using ING's infrastructure. The focus was on the focus on the maximization of data quality in the internal database and the minimization of the manual effort and the time of execution. Therefore, all the available sources were evaluated and only the ones that added significant value were incorporated. The method consists of the following five steps:

Step 1. Data Cleansing;
Step 2. Match on Cross-reference;
Step 3. Match on Legal Name;
Step 4. Match using BvD Cross-references (only for Moody's);
Step 5. Approximate String Matching Algorithm

The first four steps were implemented using SQL scripts in Microsoft Access 2010 and the fifth step using Python scripts in Portable Python 2.7.6.1. In the first step, both the internal and external data were cleansed. In the second step, the issuers were matched with the internal organizations on the existing cross-references and in the third step on the legal name. In the fourth step, which is only applicable for Moody's, the issuers were matched using Bureau van Dijk's cross references. In the last step, we used an approximate string matching algorithm on the legal name to compare the external issuers with GRID's organizations and return the GRID's organization with the highest match for each issuer. The main objectives of the step were to correctly match as many issuers as possible to GRID's organizations, to identify and exclude the issuers that do not exist in GRID in order to reduce the manual reconciliation effort, and to give accurate suggestions to the user for the issuers that should be manually reconciled to facilitate the matching procedure. Two approximate string matching algorithms were compared, the Levenshtein Distance algorithm and the Jaro-Distance algorithm.

For the validation and evaluation of our method, another data set of three different monthly files was used to avoid overfitting as much as possible. The most recent files were selected, because they would contain new issuers that did not exist in the files that were used for data analysis. The monthly files were preferred to the daily files, because the contained more records and thus had greater variety. Two types of validation indicators were used. For the first type of validation, the precision, recall and F1 score were calculated for each file and for each approximate string matching algorithm. These performance indicators are widely used in evaluating search strategies [Goutte et al. 2005; Levin et al. 2012] in order to measure the search effectiveness. The second type of validation indicator was the total number of matched issuers, the percentage of the manual effort and the total number of the issuers that do not exist in GRID.

The precision, recall and F1 show that the Levenshtein Distance algorithm is indeed better that the Jaro-Distance algorithm, even though Jaro-Distance has also high precision, recall and F1 score values. On average, the precision of the Levenshtein Distance algorithm was 93.6%, the recall was 94.47% and the F1 was 94%, while the precision of the Jaro-Distance algorithm was 85.34%, the recall was 95.83% and the F1 was 90.2%. Furthermore, the results of the new files are similarly high compared with the results of the files that were used for data analysis. This proves the validity of our method and that the phenomenon of overfitting was largely avoided. On average, 93.79% issuers were matched and the required manual effort was only 3.92%. For the reconciliation of the 3.92% of the file, our method gives the user the three highest matches for each issuer. The user can select the correct match only from these three organizations and not from the entire database. Thus, the manual effort is further reduced.

Our method is implemented and being used at ING in order to reconcile risk rating data.

7.2 Limitations

The main limitation during our research was at the available infrastructure at ING. The only available Database Management System was Microsoft Access 2010. Microsoft Access is not the most suitable DBMS to handle big amount of data. When we implemented the

approximate string matching algorithms in MS Access, the speed of execution was prohibitive. Thus, we implemented this step in Portable Python 2.7.6.1. The speed of execution in MS Access was also one of the main reasons why we did not use clusters as part of our method (Chapter 4.4.2).

In addition, as we mentioned in Chapter 4, Moody's could not provide the country of residence of the issuers that it sends to ING. Consequently, our method regarding Moody's reconciliation had to be altered and exclude equation 7 from the Python script.

7.3 Recommendations

Our experience, after researching and working on this topic, reveals that the quality of the internal and external data is not very high. For the improvement of the quality of the internal data, first of all, all the duplicate records should be identified and removed prior to the reconciliation. This activity is known as deduplication and the fifth step of our method can be used for this purpose. The second way for facilitating the reconciliation process and therefore improving the quality of the internal data is to set requirements for onboarding a vendor. In our opinion, the following requirements should be met:

- The vendor should provide at least the following attributes for each issuer:
    - Unique ID;
    - Statutory legal name;
    - Country of residence;
    - Town of residence;
    - Customer Type.
- The records of the issuer file should not contain any duplicates
- The statutory legal name of the issuers should not contain any other words, such as NEW, apart from the name itself.

7.4 Future Work

Although our method has very high precision, recall and F1 and it matches on average 93.79% issuers, it can always be improved. First of all, we only used the statutory legal name and the country of residence of an issuer for matching, because Moody's, Fitch and S&P did not provide any other attributes that describe an issuer. Therefore, whenever additional information is provided, it should be used in the reconciliation process in order to improve the accuracy of the method and correctly match more issuers.

Apart from the design of our method, the implementation can be improved. In our case, we had several limitations on the available database management systems. The only available system was Microsoft Access 2010. This system cannot process quickly big amount of data and thus, if our method was implemented in another database management system, the time of execution would have been significantly reduced. In addition, we would not have to implement the fifth step in Python scripts and the whole method would have been implemented in one system and not in two. Consequently, the manual effort for exporting and importing the files from one system to the other would have been reduced and the user's work would have been simplified.

Appendix A

In this section, details of the files, that Fitch and Moody's are providing, are given.

Fitch

Description of the attributes:

1. Report Date/Time: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the report was generated;
2. Agent Common ID: Uniquely identifies an Issuer in the IDS Database;
3. Agent CUSIP: 6-digit unique identifier of an issuer, as assigned by the Committee on Uniform Securities Identification Procedures (CUSIP);
4. Customer Identifier: The Customer Identifier field contains any identifier that was provided by a client and uploaded to their portfolio which is maintained within the IDS Database;
5. Market Sector ID: A proprietary 8-digit numeric industry classification code of the rated entity;
6. Country Name: The Name of the Country as it appears in the Fitch Ratings Database;
7. Issuer ID: Unique internal Fitch issuer identifier. This proprietary numeric code is permanently assigned to each record, and will never be reused in conjunction with any other issuer;
8. Issuer Name: Full registered name of the Issuer;
9. Issuer Record Change Code Date/Time: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the last change was made to the record;
10. Country Code: Nation of legal registration or domicile of the issuer, typically expressed according to the ISO3166-1 Alpha-3 country code abbreviation standard;
11. State/Province: State of legal registration or legal domicile of issuer, where available;
12. Issuer Currency Code: The currency in which an Issuer's securities or bonds are issued, typically expressed according to the ISO 4217 Alpha-3 currency code abbreviation standard;
13. Record Group Type Code: Proprietary internal classification of the record that defines its primary business role, as assigned by Fitch Ratings
14. Dow Jones Ticker: Stock market identifier as designated by Dow Jones and Company;
15. NAIC Industry Company Identifier: 5-digit unique identifier of an insurance company, as assigned by the National Association of Insurance Commissioners (NAIC)
16. SIC Code: 4-digit numeric Standard Industrial Classification (SIC) code as assigned by the U.S. Government that designates the industry of the issuer
17. ICB Group/Super-Sector Code: 4-digit numeric Industry Classification Benchmark code as developed by Dow Jones and FTSE;
18. NAICS Industry Code: A 5- or 6-digit numeric industry identifier, the North American Industry Classification System Code was established by the U.S. Office of Management and Budget in conjunction with statistical agencies of Canada and Mexico. The standard was adopted in 1997 to replace the SIC system;
19. Long-Term Issuer Default Rating: A Long-Term Issuer Default Rating (LT IDR) measures the probability that an issuer would default on its outstanding debt obligations with a time horizon of greater than 12 months for most issuers;
20. LT IDR Action: Last relevant activity of the associated LT IDR rating;
21. LT IDR Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT IDR rating took effect;

22. LT IDR Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months
23. Long-Term Issuer Rating: A Long-Term rating is an evaluation of credit risk and the projected capacity for timely payment of financial commitments. These have a time horizon of greater than 12 months for most issuers;
24. LT Issuer Rating Action: Last relevant activity of the associated LT rating;
25. LT Issuer Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT rating took effect;
26. LT Issuer Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
27. Long-Term National Issuer Rating: National Ratings are an assessment of credit quality relative to the rating of the "best" credit risk in a country. This "best" risk will normally be assigned to all financial commitments issued or guaranteed by the sovereign state. National Ratings are not intended to be internationally comparable. LT National Ratings typically have a time horizon of greater than 12 months;
28. LT National Issuer Rating Action: Last relevant activity of the associated LT National Rating;
29. LT National Issuer Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT National Rating took effect;
30. LT National Issuer Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
31. Long-Term Local Currency Issuer Default Rating: Local Currency IDR Ratings measure the probability that an issuer would default on its outstanding debt obligations in the currency of the locality in which it is domiciled and does not account for situations where it would be impossible to convert local currency into foreign currency or make transfers between sovereign jurisdictions. LT ratings typically have a time horizon of greater than 12 months;
32. LT Local Currency IDR Action: Last relevant activity of the associated LT Local Currency IDR rating;
33. LT Local Currency IDR Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT Local Currency IDR rating took effect;
34. LT Local Currency IDR Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
35. Long-Term Local Currency Issuer Rating: Local Currency credit ratings measure the likelihood of repayment in the currency of the locality in which the issuer is domiciled and does not account for situations where it would be impossible to convert local currency into foreign currency or make transfers between sovereign jurisdictions. LT ratings typically have a time horizon of greater than 12 months;
36. LT Local Currency Rating Action: Last relevant activity of the associated LT Local Currency rating;
37. LT Local Currency Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT Local Currency rating took effect;

38. LT Local Currency Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
39. Short-Term Issuer Default Rating: A Short-Term Issuer Default Rating measures the probability that an issuer would default on its outstanding debt obligations with a time horizon of less than 13 months for most issuers;
40. ST IDR Action: Last relevant activity of the associated ST IDR rating;
41. ST IDR Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated ST IDR rating took effect;
42. ST IDR Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
43. Short-Term Issuer Rating: A Short-Term rating is an evaluation of credit risk and the projected capacity for timely payment of financial commitments. These have a time horizon of less than 13 months for most issuers;
44. ST Issuer Rating Action: Last relevant activity of the associated ST rating;
45. ST Issuer Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated ST rating took effect;
46. ST Issuer Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
47. Short-Term National Issuer Rating: National Ratings are an assessment of credit quality relative to the rating of the "best" credit risk in a country. This "best" risk will normally be assigned to all financial commitments issued or guaranteed by the sovereign state. National Ratings are not intended to be internationally comparable. ST National Ratings typically have a time horizon of less than 13 months;
48. ST National Issuer Rating Action: Last relevant activity of the associated ST National rating;
49. ST National Issuer Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated ST National rating took effect;
50. ST National Issuer Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
51. Short-Term Local Currency Issuer Default Rating: : Local Currency IDR Ratings measure the probability that an issuer would default on its outstanding debt obligations in the currency of the locality in which it is domiciled and does not account for situations where it would be impossible to convert local currency into foreign currency or make transfers between sovereign jurisdictions. ST ratings typically have a time horizon of less than 13 months;
52. ST Local Currency IDR Action: Last relevant activity of the associated ST Local Currency IDR rating;
53. ST Local Currency IDR Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated ST Local Currency IDR rating took effect;
54. ST Local Currency IDR Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;

55. Short-Term Local Currency Issuer Rating: Local Currency credit ratings measure the likelihood of repayment in the currency of the locality in which the issuer is domiciled and does not account for situations where it would be impossible to convert local currency into foreign currency or make transfers between sovereign jurisdictions. ST ratings typically have a time horizon of less than 13 months;
56. ST Local Currency Issuer Rating Action: Last relevant activity of the associated ST Local Currency rating;
57. ST Local Currency Issuer Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated ST Local Currency rating took effect;
58. ST Local Currency Issuer Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
59. Bank Individual Rating: An assessment of how a bank would be viewed if it were entirely independent and could not rely on external support. These ratings are designed to assess an entity's exposure to and management of risk, and represent the likelihood that it would run into significant difficulties in which it would require assistance;
60. Bank Individual Rating Action: Last relevant activity of the associated Individual rating;
61. Bank Individual Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated Individual rating took effect;
62. Bank Individual Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
63. Bank Support Rating: Support ratings assess the likelihood that the entity would receive external assistance and financial support in cases of extreme hardship or default;
64. Bank Support Rating Action: Last relevant activity of the associated Support rating;
65. Bank Support Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated Support rating took effect;
66. Bank Support Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating; however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
67. Insurer Financial Strength Rating: Evaluation of an insurance company's ability to repay on indemnities and other remuneration obligations in accord with the terms of the original policy;
68. IFS Rating Action: Last relevant activity of the associated IFS rating;
69. IFS Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated IFS rating took effect;
70. IFS Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
71. Long-Term National Insurer Financial Strength Rating: LT National IFS Ratings assess the ability of an insurer to meet policyholder and related obligations, relative to the "best" credit risk in a given country, across all industries and obligation types. These are not comparable to similar ratings of insurers in other countries;

72. LT National IFS Rating Action: Last relevant activity of the associated LT National IFS rating
73. LT National IFS Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated LT National IFS rating took effect;
74. LT National IFS Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
75. Sovereign Country Ceiling Rating: Country Ceiling Ratings gauge the risk of capital and exchange controls being imposed by the sovereign authorities that would prevent or impede the private sector's ability to convert local currency into foreign currency and transfer to non-resident creditors;
76. Sovereign Country Ceiling Rating Action: Last relevant activity of the associated Country Ceiling rating;
77. Sovereign Country Ceiling Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated Country Ceiling rating took effect;
78. Sovereign Country Ceiling Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
79. Issuer Volatility Rating: Managed Fund Volatility Ratings measure the relative sensitivity of the total return on a fund's shares to a broad array of changes in interest rates, mortgage prepayment speeds, liquidity of the portfolio, spreads, currency exchange rates, and other market conditions;
80. Issuer Volatility Rating Action: Last relevant activity of the associated Volatility rating;
81. Issuer Volatility Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated Volatility rating took effect;
82. Issuer Volatility Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating, however, Rating Watch designations are usually more immediate, typically resolved within 12 months;
83. Security Identifier Type: A switch setting to test for the existence of a CUSIP identifier. If present, refer to Field 3 for the CUSIP identifier;
84. Endorsement Compliance: A coded value denoting the regulatory status of the rating;
85. Ratings Suffix: This field identifies a Ratings Suffix that Fitch Ratings has implemented to increase transparency into what the rating addresses. Fitch recommends that these suffixes be displayed alongside the rating where applicable;
86. Viability Rating: Designed to be internationally comparable, Viability ratings (VRs) represent Fitch's view as to the intrinsic creditworthiness of an issuer. Together with the agency's support ratings framework, the VR is a key component of a bank's Issuer Default Rating (IDR);
87. Viability Rating Action: Last relevant activity of the associated Viability Rating;
88. Viability Rating Effective Date: This field is displayed in the YYYY-MM-DD HH:MM:SS format and shows when the associated Viability Rating took effect;
89. Viability Rating Alert Code: Indicates the rating's status on Rating Watch or Rating Outlook. These both assess the likely future direction of the rating; however, Rating Watch designations are usually more immediate, typically resolved within 12 months.

Moody's

In the following table, the records' structure is provided:

	Field Name	Start	Width	Type	Example
1*	Moody's Issuer Name	1	10	Numeric	826578
2*	Issuer Name	11	60	Character	WINN- Example Inc.
3	Ticker	61	15	Character	WIN
4	Long Term Rating (derived by algorithm)	76	10	Character	Ba2
5	Long Term Rating Date	86	8	Date	20050412
6	Long Term Rating Class	94	80	Character	LT Issuer Rating
7	Long Term Rating Indicator	174	5		
8	Issuer Rating	179	10	Character	Ba2
9	Issuer Rating Date	189	8	Date	20050412
10	Issuer Rating Endorsement Indicator	197	2	Character	1B
11	Issuer Rating Unsolicited Indicator	199	2	Character	3
12	Issuer Rating - Rating Office Code	201	3	Numeric	118
13	Issuer Rating Withdrawal Reason	204	8	Character	WMO
14	Issuer Rating - Foreign Currency	212	10	Character	Ba2
15	Issuer Rating - Foreign Currency Date	222	8	Date	20050412
16	Issuer Rating - FC Endorsement Indicator	230	2	Character	1B
17	Issuer Rating - FC Unsolicited	232	2	Character	3
18	Issuer Rating - FC Rating Office Code	234	3	Numeric	118
19	Issuer Rating - FC Withdrawal Reason	237	8	Character	WMO
20	Issuer Rating - Domestic Currency	245	10	Character	Ba2
21	Issuer Rating - Domestic Currency Date	255	8	Date	20050412
22	Issuer Rating - DC Endorsement Indicator	263	2	Character	1B
23	Issuer Rating - DC Unsolicited Indicator	265	2	Character	3
24	Issuer Rating – DC Rating Office Code	267	3	Numeric	118
25	Issuer Rating – DC Withdrawal Reason	270	8	Character	WMO
26	Short Term Issuer Level Rating	278	10	Character	NP
27	Short Term Issuer Leven Rating Date	288	8	Date	20050412
28	Short Term Issuer Level Rating Class	296	80	Character	Commercial Paper – Dom Curr
29	Short Term Issuer Level Rating Indicator	376	5		
30	Corporate Family Ranking	381	10	Character	Ba1

31	Corporate Family Rating Date	391	8	Date	20040625
32	Corporate Family Rating Endorsement Indicator	399	2	Character	1B
33	Corporate Family Rating Unsolicited Indicator	401	2	Character	3
34	Corporate Family Rating – Rating Office Code	403	3	Numeric	118
35	Corporate Family Rating Withdrawal Reason	406	8	Character	WMO
36	Estimated Senior Rating (derived by algorithm)	414	10	Character	Ba2
37	Estimated Senior Rating Date	424	8	Date	20040625
38	Outlook	432	15	Character	RUR
39	Outlook Date	447	8	Date	20041026
40	Watchlist Indicator	455	3	Character	ON
41	Watchlist Date	458	8	Date	20041028
42	Watchlist Reason	466	20	Character	Possible Downgrade
43	Reserved for Future Use	486	65	Character	xxx

Table A.0.1 - Moody's records structure

* Mandatory Fields

Description of attributes:

1. Moody’s Issuer Number: An identifying number assigned by Moody’s to uniquely identify each issuing company or entity;
2. Issuer Name: This field represents the issuer of the security;
3. Ticker: This field displays the Equity Ticker;
4. Long Term Rating: A complex selection algorithm that considers rating class, currency and rating date to choose which actual rating qualifies as the long term rating. The essence of the algorithm is as follows: All of the organization’s long term debt ratings are grouped into the following rating classes and ordered in the following way:
    - Issuer Level
    - Senior Unsecured
    - Subordinate
    - Preferred Stock
    - Secured

    Following this order, the most senior rating class that has an active rating is identified, and that rating is chosen. If a non-domestic rating exists within the selected rating class, it is chosen, otherwise a domestic rating is picked. If more than one rating exists within domestic or non-domestic, the best (highest) rating is picked. If more than one date exists for the best rating, the most recent rated is chosen. The algorithm excludes Management Quality, Industrial Revenue Bond, Mutual Fund and Bank Financial Strength Ratings;
5. Long Term Rating Date: It is displayed in YYYYMMDD format and shows the date of the Long Term Rating;
6. Long Term Rating Class: This field displays the rating class of the Long Term Rating;

7. Long Term Rating Indicator: The Long Term Rating Indicator will be set to '(hyb)' in the event that all of the ratings for the Rating Class that was selected for the Long Term Issuer Rating are for Hybrid instruments. It can also be blank;
8. Issuer Rating: It is an opinion of the ability of an entity to honor financial obligations and contracts;
9. Issuer Rating Date: It is displayed on the YYYYMMDD format and shows the date Moody's assigned the Issuer rating;
10. Issuer Rating Endorsement Indicator: This field contains a value that represents the endorsements for the Issuer Rating. For example, the value "1B" indicates that the rating is EU Endorsed.

Value	Description
1A	The rating is EU Rated
1B	The rating is EU Endorsed
1C	The rating is EU Qualified by Extension
Blank	The rating is neither an EU Rated nor an EU Endorsed rating

11. Issuer Rating Unsolicited Indicator: Moody's identifies whether ratings are solicited or unsolicited in our data feed products by using a 2 character Unsolicited Indicator. The potential values of the Unsolicited Indicator are listed in the table below:

Un-solicited Indicator Code	Description
(blank)	Solicited and Participating
01	Non-Participating
02	Unsolicited (Global)
03	Non-Participating/Unsolicited (Global)
04	Unsolicited (EU)
05	Non-Participating/ Unsolicited (EU)
08	Unsolicited (Japan)
09	Non-participating/ Unsolicited (Japan)

12. Issuer Rating – Rating Office Code: This field contains a 3 character value that indicates the Moody's office where the Issuer Rating was assigned. In order to identify the office a mapping table is provided;
13. Issuer Rating Withdrawal Reason Code: This field contains a value indicating the reason that a rating was withdrawn;

Withdrawal Reason Code	Withdrawal Reason
WMO	Obligation is not outstanding
WMR	Reorganization
WMI	Inadequate information
WML	Bankruptcy/Liquidation/Debt restructuring
WMS	Business Reasons
WMC	Conflict of interest
WMB	Clerical error (subset of Business Reasons)
WMQ	Regulatory Requirements
WMF	Small pool factor

(blank)	No withdrawal reason

14. Issuer Rating - Foreign Currency: Issuer Rating in foreign currency is an opinion of the ability of an entity to honor financial obligation and contracts denominated in foreign currency. It is subject to Moody's Foreign Currency Country Ceiling;
15. Issuer Rating - Foreign Currency Date: This field is displayed in YYYYMMDD format and shows the date Moody's assigned the Issuer Rating;
16. Issuer Rating - FC Endorsement Indicator: This field contains the same kind of information as the Issuer Rating Endorsement Indicator field;
17. Issuer Rating - FC Unsolicited Indicator: This field contains the same kind of information as the Issuer Rating Unsolicited Indicator field;
18. Issuer Rating - FC Rating Office Code: This field contains the same kind of information as the Issuer Rating – Rating Office Code field;
19. Issuer Rating - FC Withdrawal Reason Code: This field contains the same kind of information as the Issuer Rating Withdrawal Reason Code field;
20. Issuer Rating - Domestic Currency: This is an opinion of the ability of entity to honor financial obligation and contracts denominated in domestic currency;
21. Issuer Rating - Domestic Currency Date: This field is displayed in YYYYMMDD format and shows the date Moody's assigned the Issuer Rating;
22. Issuer Rating - DC Endorsement Indicator: This field contains the same kind of information as the Issuer Rating Endorsement Indicator field;
23. Issuer Rating - DC Unsolicited Indicator: This field contains the same kind of information as the Issuer Rating Unsolicited Indicator field;
24. Issuer Rating - DC Unsolicited Indicator: This field contains the same kind of information as the Issuer Rating Unsolicited Indicator field;
25. Issuer Rating - DC Rating Office Code: This field contains the same kind of information as the Issuer Rating - Rating Office Code field;
26. Issuer Rating - DC Withdrawal Reason Code: This field contains the same kind of information as the Issuer Rating Withdrawal Reason Code field;
27. Short Term Issuer Level Rating: This field uses a rating class precedence logic to pick the most relevant short-term rating. The logic follows the following rules:
    - Considers definitive ratings before classes with prospective ratings
    - Considers short term rating class precedence (unsecured is considered before backed
    - Considers non-domestic before domestic

    If the selected rating classes have multiple ratings (same currency), the highest rating is picked. If both ratings are the same, the one with the most recent date is picked;
28. Short Term Issuer Level Rating Date: This field is displayed in the YYYYMMDD format and shows the date Moody's assigned the Short Term Issuer Level Recent Rating.
29. Short Term Issuer Level Rating Class: This field contains the same kind of information as the Short Term Issuer Level Recent Rating;
30. Short Term Issuer Level Rating Indicator: This field contains the same kind of information as the Long Term Rating Indicator field;

31. Corporate Family Rating (formerly Senior Implied Rating): Moody's Corporate Family Ratings are generally employed for speculative grade corporate issuers. The Corporate Family Ratings is an opinion of a corporate family's ability to honor its financial obligations and is assigned to a corporate family as if it had.
    - A single class of debt;
    - A single consolidated legal entity structure.

    The Corporate Family Rating differs from Moody's Issuer Rating, which references an obligor's senior unsecured obligations and which also reflects the obligor's actual corporate structure. By contract, the Corporate Family Raiting assumes away such structural and legal complexities;
32. Corporate Family Rating Date (formerly Senior Implied Rating Date): This field is displayed in the YYYYMMDD format and shows the date Moody's assigned the Corporate Family Rating;
33. Corporate Family Rating Endorsement Indicator: This field contains the same kind of information as the Issuer Rating Endorsement Indicator field;
34. Corporate Family Unsolicited Indicator: This field contains the same kind of information as the Issuer Rating Unsolicited Indicator field;
35. Corporate Family Rating - Rating Office Code: This field contains the same kind of information as the Issuer Rating - Rating Office Code field;
36. Corporate Family Rating - Withdrawal Reason Code: This field contains the same kind of information as the Issuer Rating Withdrawal Reason Code field;
37. Estimated Senior Rating: This field is derived algorithmically from ratings assigned to an issuer's other rated debt via a simple notching algorithm that is intended to reflect observed ratings relationships;
38. Estimated Senior Rating Date: This field is displayed in the YYYYMMDD format and shows the date Moody's assigned the Estimated Senior Rating;
39. Outlook: This is an opinion regarding the likely direction of a rating over the medium term. Where assigned, rating outlooks fall into the following four categories: Positive (POS), Negative (NEG), Stable (STA) and Developing (DEV – contingent upon an event). In the few instances where an issuer has multiple outlooks of differing directions, an "(m)" modifier will be displayed, and Moody's written research will describe any differences and provide the rationale for these differences. The RUR (Rating(s) Under Review) designation indicates that the issuer has one or more ratings under review for possible change, and thus overrides the outlook designation. When an outlook has not been assigned to an eligible entity, NOO (No Outlook) may be displayed. RWR indicates withdrawn ratings;
40. Outlook Date: This field is displayed in the YYYYMMDD format and shows he effective date for the Outlook;
41. Watchlist Indicator: This field indicates whether Moody's has taken a "Watchlist Action" on any one of the rated debts for the selected issuer. This implies that Moody's is actively considering or has previously considered changing or confirming the current rating. A change to the Watchlist Indicator field will cause a record to be transmitted. The possible valued of this field are the following:

Value	Description
CFO	Confirm Only (Rating Confirmation)

OFF	Taken Off Watch
ON	Placed On Watch

42. Watchlist Date: This field is displayed in the YYYYMMDD format and indicates when the Watchlist Action took place. If multiple issuers are on watch, this field will be blank;
43. Watchlist Reason: For securities “On Watch”, Moody’s will indicate the more-likely direction of a rating change (upgrade, downgrade or uncertain). Different Watchlist Reason codes are used depending upon the status of the Watchlist Indicator field. The possible values are:
    - Possible Upgrade
    - Possible Downgrade
    - Uncertain - on watch for possible upgrade or downgrade
    - Multiple - indicates multiple debts of the issuer are on watch

Appendix B

Total Records	621	621	621	621	621	621	621	621	621
Lower Threshold	**79**	**80**	**81**	**82**	**83**	**84**	**85**	**79**	**80**
Upper Threshold	**95**	**95**	**95**	**95**	**95**	**95**	**95**	**96**	**96**
Records	161	171	197	219	251	280	308	161	171
Correct	160	170	193	214	246	272	295	160	170
Correct %	99,38%	99,42%	97,97%	97,72%	98,01%	97,14%	95,78%	99,38%	99,42%
False (Exist)	1	1	3	4	4	7	12	1	1
False (Exist) %	0,62%	0,58%	1,52%	1,83%	1,59%	2,50%	3,90%	0,62%	0,58%
False (Wrong match)	0	0	1	1	1	1	1	0	0
False (Wrong match) %	0,00%	0,00%	0,51%	0,46%	0,40%	0,36%	0,32%	0,00%	0,00%
False	1	1	4	5	5	8	13	1	1
False %	0,62%	0,58%	2,03%	2,28%	1,99%	2,86%	4,22%	0,62%	0,58%
Records	435	425	399	377	345	316	288	442	432
Correct	78	78	76	75	75	81	66	84	84
Correct %	17,93%	18,35%	19,05%	19,89%	21,74%	25,63%	22,92%	19,00%	19,44%
False (New)	348	338	315	294	262	237	214	349	339
False (New) %	80,00%	79,53%	78,95%	77,98%	75,94%	75,00%	74,31%	78,96%	78,47%
False (Wrong match)	9	9	8	8	8	8	8	9	9
False (Wrong match) %	2,07%	2,12%	2,01%	2,12%	2,32%	2,53%	2,78%	2,04%	2,08%
Records	25	25	25	25	25	25	25	18	18
Correct	21	21	21	21	21	21	21	15	15
Correct %	84,00%	84,00%	84,00%	84,00%	84,00%	84,00%	84,00%	83,33%	83,33%
False (New)	4	4	4	4	4	4	4	3	3
False (Exist) %	16,00%	16,00%	16,00%	16,00%	16,00%	16,00%	16,00%	16,67%	16,67%
False (Wrong match)	0	0	0	0	0	0	0	0	0
False (Wrong match) %	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
False	4	4	4	4	4	4	4	3	3
False %	16,00%	16,00%	16,00%	16,00%	16,00%	16,00%	16,00%	16,67%	16,67%
Total	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**
Correct	**185**	**195**	**218**	**239**	**271**	**297**	**320**	**178**	**188**
Correct %	**29,79%**	**31,40%**	**35,10%**	**38,49%**	**43,64%**	**47,83%**	**51,53%**	**28,66%**	**30,27%**
False	**5**	**5**	**8**	**9**	**9**	**12**	**17**	**4**	**4**
False %	**0,81%**	**0,81%**	**1,29%**	**1,45%**	**1,45%**	**1,93%**	**2,74%**	**0,64%**	**0,64%**
Manual Effort	**435**	**425**	**399**	**377**	**345**	**326**	**288**	**442**	**432**
Manual Effort %	**70,05%**	**68,44%**	**64,25%**	**60,71%**	**55,56%**	**52,50%**	**46,38%**	**71,18%**	**69,57%**

621	621	621	621	621	621	621	621	621	621	621	621
81	**82**	**83**	**84**	**85**	**79**	**80**	**81**	**82**	**83**	**84**	**85**
96	**96**	**96**	**96**	**96**	**97**	**97**	**97**	**97**	**97**	**97**	**97**
197	219	251	280	308	161	171	197	219	251	280	308
193	214	246	272	295	160	170	193	214	246	272	295
97,97%	97,72%	98,01%	97,14%	95,78%	99,38%	99,42%	97,97%	97,72%	98,01%	97,14%	95,78%
3	4	4	7	12	1	1	3	4	4	7	12
1,52%	1,83%	1,59%	2,50%	3,90%	0,62%	0,58%	1,52%	1,83%	1,59%	2,50%	3,90%
1	1	1	1	1	0	0	1	1	1	1	1
0,51%	0,46%	0,40%	0,36%	0,32%	0,00%	0,00%	0,51%	0,46%	0,40%	0,36%	0,32%
4	5	5	8	13	1	1	4	5	5	8	13
2,03%	2,28%	1,99%	2,86%	4,22%	0,62%	0,58%	2,03%	2,28%	1,99%	2,86%	4,22%
406	384	352	323	295	455	445	419	397	365	336	308
82	81	81	77	72	94	94	92	91	91	87	82
20,20%	21,09%	23,01%	23,84%	24,41%	20,66%	21,12%	21,96%	22,92%	24,93%	25,89%	26,62%
316	295	263	238	215	352	342	319	298	266	241	218
77,83%	76,82%	74,72%	73,68%	72,88%	77,36%	76,85%	76,13%	75,06%	72,88%	71,73%	70,78%
8	8	8	8	8	9	9	8	8	8	8	8
1,97%	2,08%	2,27%	2,48%	2,71%	1,98%	2,02%	1,91%	2,02%	2,19%	2,38%	2,60%
406	384	352	323	295	455	445	419	397	365	336	308
18	18	18	18	18	5	5	5	5	5	5	5
15	15	15	15	15	5	5	5	5	5	5	5
83,33%	83,33%	83,33%	83,33%	83,33%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
3	3	3	3	3	0	0	0	0	0	0	0
16,67%	16,67%	16,67%	16,67%	16,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
0	0	0	0	0	0	0	0	0	0	0	0
0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
3	3	3	3	3	0	0	0	0	0	0	0
16,67%	16,67%	16,67%	16,67%	16,67%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
621	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**	**621**
211	**232**	**264**	**290**	**313**	**165**	**175**	**198**	**219**	**251**	**277**	**300**
33,98%	**37,36%**	**42,51%**	**46,70%**	**50,40%**	**26,57%**	**28,18%**	**31,88%**	**35,27%**	**40,42%**	**44,61%**	**48,31%**
7	**8**	**8**	**11**	**16**	**1**	**1**	**4**	**5**	**5**	**8**	**13**
1,13%	**1,29%**	**1,29%**	**1,77%**	**2,58%**	**0,16%**	**0,16%**	**0,64%**	**0,81%**	**0,81%**	**1,29%**	**2,09%**
406	**384**	**352**	**323**	**295**	**455**	**445**	**419**	**397**	**365**	**336**	**308**
65,38%	**61,84%**	**56,68%**	**52,01%**	**47,50%**	**73,27%**	**71,66%**	**67,47%**	**63,93%**	**58,78%**	**54,11%**	**49,60%**

Table A.0.1 - Thresholds for S&P without Country of Residence

Appendix C

 # PYTHON SCRIPT USING LEVENSHTEIN DISTANCE ALGORITHM AND INCLUDING THE COUNTRY OF RESIDENCE

```
1 import time
2 import sys
3 import csv
4 import io
5 import Levenshtein as ls
6 import datetime
7
8 Inginput = 'GRID.txt'
9 extinput = 'Issuer.txt'
10 outputLow = 'outputc_Issuer_New '+str(datetime.datetime.now().strftime("%Y-%m-%d %H-%M-%S"))+'.txt'
11 outputMiddle = 'outputc_Issuer_Manual '+str(datetime.datetime.now().strftime("%Y-%m-%d %H-%M-%S"))+'.txt'
12 outputHigh = 'outputc_Issuer_Matched '+str(datetime.datetime.now().strftime("%Y-%m-%d %H-%M-%S"))+'.txt'
11
13 # Register the start time
14 Start = time.time()
15
16 with open(inginput, 'rb') as csving, open(extinput, 'rb') as csvext:
17     filereaderING = list(csv.reader(csving,delimiter=';'))
18     filereaderEXT = list(csv.reader(csvext,delimiter=';'))
19     bufsize = 0
20     f = open(outputLow, 'w', bufsize)
21     g = open(outputMiddle, 'w', bufsize)
22     h = open(outputHigh, 'w', bufsize)
23
24     try:
25         for rowext in filereaderEXT:
26             # initialize the variables
27             maxID1 = '0'
28             maxName1 = '0'
29             maxRatio1 = 0
30             maxCountry1 = '0'
31             maxID2 = '0'
32             maxName2 = '0'
33             maxRatio2 = 0
34             maxCountry2 = '0'
35             maxID3 = '0'
36             maxName3 = '0'
37             maxRatio3 = 0
38             maxCountry3 = '0'
39             for row in filereaderING:
40                 ratio = 0
41                 # get the levenhstein ratio
42                 ratio = ls.ratio(row[1].lower(),rowext[1].lower())
43
44                 if row[2] != rowext[2]:
45                     ratio = 0.9 * ratio
46
47                 # check if the new match is higher than the previous and select the highest
48                 if ratio > maxRatio3:
49                     maxID1 = maxID2
50                     maxRatio1 = maxRatio2
```

```
51                              maxName1 = maxName2
52                              maxCountry1 = maxCountry2
53
54                              maxID2 = maxID3
55                              maxRatio2 = maxRatio3
56                              maxName2 = maxName3
57                              maxCountry2 = maxCountry3
58
59                              maxID3 = row[0]
60                              maxRatio3 = ratio
61                              maxName3 = row[1]
62                              maxCountry3 = row[2]
63                      elif ratio > maxRatio2:
64                              maxID1 = maxID2
65                              maxRatio1 = maxRatio2
66                              maxName1 = maxName2
67                              maxCountry1 = maxCountry2
68
69                              maxID2 = row[0]
70                              maxRatio2 = ratio
71                              maxName2 = row[1]
72                              maxCountry2 = row[2]
73
74                      elif ratio > maxRatio1
75                              maxID1 = row[0]
76                              maxRatio1 = ratio
77                              maxName1 = row[1]
78                              maxCountry1 = row[2]
79
80              if int(maxRatio3*100) < 80:
81                      f.write("%s;%s;%s;%s;%s;%s;%s\n"                                        %
(rowext[0],rowext[1],maxName3,maxID3,rowext[2],maxCountry3,str(int(maxRatio3*100))))
82              elif int(maxRatio3*100) > 96:
83                      h.write("%s;%s;%s;%s;%s;%s;%s\n"                                        %
(rowext[0],rowext[1],maxName3,maxID3,rowext[2],maxCountry3,str(int(maxRatio3*100))))
84              else:
85                      g.write("%s;%s;%s;%s;%s;%s;%s\n"                                        %
(rowext[0],rowext[1],maxName3,maxID3,rowext[2],maxCountry3,str(int(maxRatio3*100))))
86                      g.write("%s;%s;%s;%s;%s;%s;%s\n"                                        %
(rowext[0],rowext[1],maxName2,maxID2,rowext[2],maxCountry2,str(int(maxRatio2*100))))
87                      g.write("%s;%s;%s;%s;%s;%s;%s\n"                                        %
(rowext[0],rowext[1],maxName1,maxID1,rowext[2],maxCountry1,str(int(maxRatio1*100))))
88
89      except csv.Error as e:
90              sys.exit('error %s' % (e))
91
92 f.close()
93 g.close()
94 h.close()
95
96 end = time.time()
97 print 'Duration : %s' % (end - start)
```

References

[1] Altman, Edward I., and Herbert A. Rijken. "How rating agencies achieve rating stability." *Journal of Banking & Finance* 28.11 (2004): 2679-2714.

[2] Basel committee on banking supervision, “Basel III: A global regulatory framework for more resilient banks and banking systems”, Bank for international settlements, December 2010

[3] Basel committee on banking supervision, “Basel III: A global regulatory framework for more resilient banks and banking systems”, Bank for international settlements, December 2010

[4] Basel committee on banking supervision, “Progress report on implementation of the Basel regulatory framework”, Basel II, Bank for international settlements, April 2014

[5] Basel committee on banking supervision, “The Internal Ratings-Based Approach”, consultative document, January 2001

[6] Bizer, Christian, et al. "The meaningful use of big data: four perspectives--four challenges." *ACM SIGMOD Record* 40.4 (2012): 56-60.

[7] Black, Paul E., "Jaro-Winkler", Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Technology, http://xlinux.nist.gov/dads//HTML/jaroWinkler.html

[8] Black, Paul E., "Levenshtein distance", Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Technology, 2014, http://xlinux.nist.gov/dads//HTML/Levenshtein.html

[9] Bouzeghoub, Mokrane, et al. "Heterogeneous data source integration and evolution." *Database and Expert Systems Applications*. Springer Berlin Heidelberg, 2002.

[10] Cantor, R. (2001): Moody’s investors service response to the consultative paper issued by the Basel Committee on Banking Supervision and its implications for the rating agency industry. Journal of Banking and Finance 25, 171-186.

[11] Caruso, Francesco, et al. "Telcordia's database reconciliation and data quality analysis tool." *VLDB*. 2000.

[12] Charras, Christian, and Thierry Lecroq. *Handbook of exact string matching algorithms*. King's College Publications, 2004.

[13] Chen, Junghui, Yungchih Peng, and Jose Co Munoz. "Correntropy estimator for data reconciliation." *Chemical Engineering Science* 104 (2013): 1019-1027.

[14] Crowe, Cameron M. "Data reconciliation—progress and challenges." *Journal of Process Control* 6.2 (1996): 89-98.

[15] Doan, AnHai, et al. "Object Matching for Information Integration: A Profiler-Based Approach." *IIWeb*. 2003.

[16] Dong, Xin, Alon Halevy, and Jayant Madhavan. "Reference reconciliation in complex information spaces." *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005.

[17] E. Caron, Business intelligence lecture slides, ICT in Business master programme, Leiden University, 2014

[18] Ferri, Giovanni, L-G. Liu, and Joseph E. Stiglitz. "The procyclical role of rating agencies: Evidence from the East Asian crisis." *Economic Notes* 28.3 (1999): 335-355.

[19] Ghauri, Pervez N., and Kjell Grønhaug. Research methods in business studies: A practical guide. Pearson Education, 2005.

[20] Gitman, Lawrence J., and Chad J. Zutter. *Principles of Managerial Finance 13th Edition*. Prentice Hall, 2011

[21] Goutte, Cyril, and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2005. 345-359.

[22] GRID – ING's Organization Database, INGWiki, 2014

[23] Hawkins, David, Walter J. Campbell, and Barbara A. Brown. Rating Industrial Bonds. NJ: Financial Executives Research Foundation, 1983.

[24] Investopedia, "Definition of an issuer", 2014, http://www.investopedia.com/terms/i/issuer.asp

[25] Ioannou, Ekaterini, Claudia Niederée, and Wolfgang Nejdl. "Probabilistic entity linkage for heterogeneous information spaces." *Advanced Information Systems Engineering*. Springer Berlin Heidelberg, 2008.

[26] Ioannou, Ekaterini, Nataliya Rassadko, and Yannis Velegrakis. "On generating benchmark data for entity matching." *Journal on Data Semantics* 2.1 (2013): 37-56.

[27] Jaro, Matthew A. "Probabilistic linkage of large public health data files."*Statistics in medicine* 14.5-7 (1995): 491-498.

[28] Jaro, Matthew A. "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida." Journal of the American Statistical Association 84.406 (1989): 414-420.

[29] Kuhner, Christoph. "Financial rating agencies: are they credible." *Insights into the reporting incentives of rating agencies in times of enhanced systemic risk. Schmalenbach Business Review* 53 (2001): 2-26.

[30] Lawrence, Ramon, and Ken Barker. "Integrating data sources using a standardized global dictionary." *Knowledge Discovery for Business Information Systems*. Springer US, 2002. 153-172.

[31] Lenzerini, Maurizio. "Data integration: A theoretical perspective." *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2002.

[32] Levenshtein, Vladimir I. "Binary codes capable of correcting deletions, insertions and reversals." *Soviet physics doklady*. Vol. 10. 1966.

[33] Levin, Michael, et al. "Citation-based bootstrapping for large-scale author disambiguation." Journal of the American Society for Information Science and Technology 63.5 (2012): 1030-1047.

[34] Levy, Yair, and Timothy J. Ellis. "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research." *Informing Science* 9 (2006).

[35] Löffler, Gunter. "Avoiding the rating bounce: Why rating agencies are slow to react to new information." *Journal of Economic Behavior & Organization* 56.3 (2005): 365-381.

[36] Magerman, Tom, Bart Van Looy, and Xiaoyan Song. "Data production methods for harmonized patent statistics: Patentee name harmonization." *DTEW-MSI_0605* (2006): 1-88.

[37] Magnani, Matteo, and Danilo Montesi. *A study on company name matching for database integration*. Technical Report UBLCS-07-15, 2007.
[38] Maletic, Jonathan I., and Andrian Marcus. "Data cleansing: A prelude to knowledge discovery." *Data Mining and Knowledge Discovery Handbook*. Springer US, 2010. 19-32.
[39] Moody's Inverstor Service, "Moody's rating symbols & definitions", June 2009
[40] Müller, Heiko, and Johann-Christph Freytag. *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.
[41] Naumann, Felix, Ulf Leser, and Johann Christoph Freytag. "Quality-driven integration of heterogeneous information systems." (1999).
[42] Noessner, Jan, et al. "Leveraging terminological structure for object reconciliation." *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 2010. 334-348.
[43] Özyurt, Derya B., and Ralph W. Pike. "Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes."*Computers & chemical engineering* 28.3 (2004): 381-402.
[44] Raats, B, GRID Data Structure, ING, 13 February 2013
[45] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." *IEEE Data Eng. Bull.* 23.4 (2000): 3-13.
[46] S. D. Simpson, "The banking system: Commercial banking – How banks make money", Investopedia, 2014, http://www.investopedia.com/university/banking-system/banking-system3.asp
[47] Saıs, Fatiha, Nathalie Pernelle, and Marie-Christine Rousset. "L2R: a logical method for reference reconciliation." *Proc. AAAI*. 2007.
[48] Sattler, Kai-Uwe, Stefan Conrad, and Gunter Saake. "Interactive example-driven integration and reconciliation for accessing database federations."*Information systems* 28.5 (2003): 393-414.
[49] Singhal, Yuvika, Anupama Sharma, and Ranjit Singh. "Comparison of Approaches Used for Data Reconciliation: A Survey."
[50] Spindler, A, Structural redundancy of data from wastewater treatment systems. Determination of individual balance equations, Water Research, 2014, Vol.57, pp.193-201, Web of Science
[51] The United States securities exchange act of 1934, as amended through P.L. 112-158, August 10, 2012, http://www.sec.gov/about/laws.shtml#secexact1934
[52] Thoma, Grid, et al. Harmonizing and combining large datasets–an application to firm-level patent and accounting data. No. w15851. National Bureau of Economic Research, 2010.
[53] Treacy, William F., and Mark Carey. "Credit risk rating systems at large US banks." *Journal of Banking & Finance* 24.1 (2000): 167-201.
[54] Webster, Jane, and Richard T. Watson. "Analyzing the past to prepare." *MIS quarterly* 26.2 (2002): 13-23.
[55] Wiederhold, Gio. "Mediators in the architecture of future information systems."*Computer* 25.3 (1992): 38-49.
[56] Winkler, William E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." (1990).