Analytical Credit Dataset and Data Analytics in Financial Services – Development of an Automated Data Extraction Tool for Banks and Credit Institutions

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Name: Güner Vorname: Gonca

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. Hans-Jörg von Mettenheim

Hannover, den 04. Oktober 2016

Table of Contents

1 Introduction

The recent financial crisis marked immense gaps and insufficiencies within the European banking scene. Despite the availability of a wide range of data on credit[1] and credit risk, European banks failed to identify and aggregate information on credit exposures, which was necessary for internal decision makings and also for an early detection of potentially weak institutions during the crisis period. The national regulators depended on the data for essentially needed micro-prudential supervision. Furthermore, the drastic disintegration of the financial development of the various sectors, regions, and different sizes of the companies during the turbulent period affected the situation negatively (Thun 2015a). Therefore, the European Central Bank (ECB) promotes the setup of a technological platform to collect, store, process, and disseminate analytical credit data in the European area (International Committee on Credit Reporting 2015). Thus, on 24 February 2014, the ECB ´decided on an organization of preparatory measures for the collection of granular credit data by the European System of Central Banks´ (ECB/2014/6). The decision resulted in the initiation of AnaCredit (Analytical Credit Dataset) – a granular dataset with more frequent and flexible credit and credit risk data that should fill data gaps within credit exposures and address monetary as well as micro- and macro-prudential issues. Consequently, the European System of Central Banks (ESCB) was required to acknowledge the regulation and start to initiate its implementation. The Governing Council adopted the regulation on 18 May 2016 and defined its requirements for the reporting institutions as well as the central credit registers (CCRs) and other available data sources. The adoption not only highlighted the extent and high demand on reporting processes and software, but also the challenge of relevant data collection that the reporting banks and credit institutions will face henceforth (BearingPoint 2016).

Establishing an integrated dataset on credit and credit risk data allows the ECB to monitor the credit amount and relation between reporting institutions and borrowing companies. As a result, the ECB receives information about the economic situation of the borrowing companies, the economic sectors in the European Union (EU) Member States, and the credit demand and credit supply situation. In this way, the ECB can discover

[1] The broadly defined terms *loan* and *credit* are synonymously used in the course of the present thesis.

how smaller companies are performing, how effective their monetary policy is, and which risks the banks and credit institutions are facing (Bentzien 2016).

1.1 Motivation and Relevance

Currently, the AnaCredit regulation is a strongly discussed topic within the banking and financial industry and there are several reasons that make AnaCredit a relevant, interesting and fruitful research topic. The present thesis is realized within a project of the IT Consulting Financial Services at KPMG in Frankfurt, Germany. The reasons and the motivation for investigating AnaCredit are outlined in the following.

First of all, the reporting banks and credit institutions necessarily need to make significant investments in IT- and data maintenance systems and changes to their credit processes. Furthermore, there is a need for long term flexible reporting architectures in order to meet the AnaCredit requirements. The sorting and analysis of the available data in the banks' current IT-systems and its quality is indispensable in identifying data gaps. Besides, a remarkable part of the required data has not been collected by the banks and credit institutions yet, because it was not relevant until now, so the banks and credit institutions need to find new approaches and concepts to gather the relevant data. Multiple consulting companies have identified this recent need of the banks, and hence, offer their services and support. Therefore, the present thesis not only deals with the AnaCredit regulation and its challenges, but also **creates the foundation stone of a tool development for generating the required data**.

The main purpose of the present thesis is the development of the data extraction tool GLEIF Data Retriever (GLEIF-DR) with Python Programming Language to enable banks and credit institutions to gather borrower-related data provided by the Global Legal Entity Identifier Foundation (GLEIF), where information about companies are updated daily. The purpose of the thesis translates into the following research question:

> *RQ: Why do the reporting banks and credit institutions need to develop new reporting architectures and how can the reporting banks and credit institutions extract and update borrower-related data automatically from the GLEIF website to meet the AnaCredit requirements?*

5 Conclusion and Outlook

RQ: 'Why do the reporting banks and credit institutions need to develop new reporting architectures and how can the reporting banks and credit institutions extract and update borrower-related data automatically from the GLEIF website to meet the AnaCredit requirements?'

The European banking scene proved remarkable gaps and insufficiencies within the recent financial crisis, although a wide range of credit and credit risk data was available. Besides various factors that were involved, European banks failed to identify potentially weak institutions during this turbulent period. Consequently, the ECB established the AnaCredit regulation in order to create a granular dataset to collect, store, process, and disseminate analytical credit and credit risk data in the European area. The new regulation affects bank loans to legal entities exceeding a minimum threshold of 25,000 Euro in the European area. The dataset consists of 95 attributes in total per loan and imposes high requirements on the approximately 5,000 banks and credit institutions, which are subject to the new regulation. As this thesis showed, the reporting banks and credit institutions need to develop new reporting architectures in order to fill the data gaps, because a large part of the required data is missing in their current IT- and data maintenance systems, while 15 million borrowers and about 100 million credits are affected by AnaCredit.

For this reason, the automated data extraction tool GLEIF-DR has been developed with Python Programming Language, which solves the problem of retrieving relevant data, in fact, 7 of the Counterparty Reference Data attributes from the GLEIF Concatenated XML File. The GLEIF-DR can be used by the reporting banks and credit institutions to retrieve up-to-date information about LEI-registered companies, which is provided by the GLEIF. In comparison to GLEIF's LEI Search Tool, the GLEIF-DR is able to retrieve information about multiple companies and their subsidiaries at once due to the implemented Fuzzy String Matching. The program is ready to be deployed and is used within the AnaCredit project at KPMG; however, there is improvement potential for the future regarding the long runtime of the code and extensions concerning the retrieve of data from PDFs and other file formats. Moreover, it is recommended to implement the

program in the reporting institutions current IT- and data maintenance systems in order to ensure smooth operations within one integrated program, instead of working with multiple smaller programs. The use of the GLEIF-DR saves time and personnel costs, while lowering the probability of potential errors that could occur during the manual processing.

The GLEIF-DR represents the foundation stone for an advanced software solution and is a tool that handles XML scraping, while further extensions are recommendable, such as retrieving information from other file formats like PDF, and other various formats. This practice requires profound machine learning skills and expert knowledge and has not been handled within this research. Extracting data from PDFs and other file formats for the AnaCredit project constitutes a research field for extended analysis based on the present thesis.