A Data Integration System for a National Register

**Keywords:** record linkage, machine learning, automated data processing

# Introduction

Our organisation is operating a register which contains reference data on institutional units following the definition given by the corresponding guideline (legal units, branches, investment funds). It facilitates the integration of several internal and external databases as well as primary reporting’s through the allocation of different identifiers for the reporting units. It enables a high flexibility with regards to analysis; hence the collected data can be used for statistical and non-statistical purposes, both across institutions and across different user groups within institutions.

The data input is characterised by a multi-source approach. There are over 1.700 reporting agents and currently already three internal registers that are continuously integrated. These data sources differ significantly in their quality and data structures. In the near future further commercial and official data sources will be integrated.

There is no stable national identifier with full coverage available to describe the respective reporting units. The aim of the integration system of the national register is to classify (“match”/ “non-match”) and consolidate every received reporting unit on a highly automated basis, since the data throughput is too high to deal with it manually. To fulfil the formulated requirements of the corresponding regulation and guideline, the produced data quality needs to be very high. Thus the implemented algorithms are precision oriented.

In our Poster we aim to describe the methods and workflows which are used for the data integration as a part of the transactional system to achieve a high quality consolidated database. The focus lies on the automated integration of units with unstable identifiers.

# Methods

The data sources that are integrated in the database have both lexical and structural heterogeneity. [1] Mappings, extractions and concatenations are implemented to deal with the structural heterogeneity. For the lexical heterogeneity there is an expression language interface to define the pre-processing rules. Especially the string attributes need to be pre-processed to be useful for the record linkage process. Information retrieval methods are applied (f. e. stemming) to reduce the “noise” in the strings.

For the record linkage of the entities there are different deterministic stages which are executed hierarchically. Record linkage describes the identification of the same entities which might possess different attributes to describe them. [2] Pairs of data are combined over identifiers as well as over soft attributes like the standardised name. Every pair is validated over more complex (more expensive) metrics, like string distances. The deterministic stages generate “quick wins” on a high quality level, with the price of a low classification rate. The algorithm takes into account the time slices of the entities and their attributes as well as the clerical interaction. Transitive closure and data integrity have to be kept in mind through the parallel processing both manually and automated. The last stage of the record linkage is so far implemented as a prototype. It comprises a probabilistic record linkage which is based on a supervised machine learning (random forest) classification model. It balances the low classification rate of the deterministic stages with the price of high transaction costs.

For the highly automated block-compounding algorithm, veracity, velocity and modus of the reported attribute are taken into account. The blocks reflect both logical links as well as given correlations. By accepting block-attributes from lower-priority sources, in case of agreement with partial blocks from higher-priority sources, completeness is promoted. Some attributes, like the identifiers might be overlooked clerically to assure a high quality of attributes.

# Results

Since February 2018 the transactional system is processing data from the reporting agents and from one internal database. The system is operating with a pre-processing layer which processes lexical and structural heterogeneity. The record linkage is implemented in several deterministic stages, as well as the automatic block compounding.

The register was mainly set up in the last months, therefore the results, like the ratios, are still shifting. There was a large number of new entries in the first month, (≈57 % February-May) which is decreasing constantly (≈33 % May-September). The ratio in May 2018 for automatically linked units was quite low with ≈24 %. Nearly ≈19 % of the reporting units were not classified at that point. One reason for a non-classification was a failed verification of a data pair (≈8%), f. e. one reporting unit has two possible candidates (possible transitive closure). Another reason for a non-classification was that the mandatory criteria for a new entry were not met and no possible candidate was found (≈11%). The quality of the classification is very high. A sample of the classified entities was manually checked for false classification. Standard measures like accuracy (≈0,99) show very high results for the quality of the matches and non matches.

With the developed probabilistic record linkage prototype the unclassified entities can be reduced by ≈48 %.

All linked reporting units are consolidated with a block compounding algorithm which uses veracity, velocity and the modus of the attribute value.

# Conclusions

The quality of the implemented pre-processing, record linkage stages and compounding algorithm is satisfying. As the criteria’s for matching with already existing entries or for creating new entries are strict, there are nearly no false classifications. This goal was prioritized regarding the given requirements described in the regulation and guideline. On the other hand the overall classification rate is too low and needs improvement. The random forest classification model which is used as a prototype is hereby promising to reduce the number of unclassified data enormously.

Since the register is still in an early phase it is missing historical data from which could be learnt. F.e. the compounding algorithm still needs to “learn” about the veracity of the reported attributes. The precision of quality measurements of the compounding algorithm will increase over time by adding data.

# References

1. A. K. Elmagarmid, P. G. Ipeirotis and V.S. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering (2007), 1-16.
2. P. Christen, Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection (2012) Springer-Verlag: Berlin Heidelberg.