# Current Status Master Thesis

## Base Matcher research:

* Analyzed papers and current successful OAEI Participants, especially YAM++
* Considering for now 24 Matcher
  + Hamming distance
  + Jaro Winkler
  + Jaro
  + Levenshtein
  + Needleman Wunsch
  + Ngram Distance
  + SMOA / STOILOIS / ISUB
  + Least Common Substring Distance
  + String equality
  + Prefix
  + Suffix
  + Monge Elkan
  + Jiang Conrath
  + Lin
  + Wu Palmer
  + TFIDF + cosine
  + Soft TFIDF + jaro
  + Jaro TFIDF
  + Jaccard
  + Level 2 jaro winkler
  + Level 2 Monge Elkan
  + TFIDF + Cosine, on comments, labels and data properties
* Furthermore implemented the following preprocessing techniques
  + Stemming
  + underscore, camel case tokenization

## Proof-of-Concept Pipeline:

* A Scala program runs all base matchers, wrapped in Alignment API matcher and saves the output to a csv file containing a similarity score of the base matcher for the mentioned matching relations
* In parallel for each base matcher the optimal threshold is computed and the best result in terms of precision, recall and f-measure is stored
* Now the meta matcher is triggered, he got as an input the computed similarity vector and performs the following steps
  + Reduce the features which correlate
  + Perform and Clustering based Outlier Detection
  + Compute the Cluster-base Outlier Factor
  + Select positive outliers
  + Normalize the outlier score to a scale 0 to 1
  + Use this factor as the outlier score
  + Optimize the threshold
    - TODO Check if there are some rules of thumb for a good threshold
* After all datasets haven been matched, compute the following two baselines:
  + Best average performing Base Matcher, based on Precision, Recall and F-Measure
  + The average of the best performing base matcher for each dataset (Average of Precision, Recall, Fe-Measure)
  + REMARK: Currently not based on the aggregated TP, FP ,FN but on the average of P,R,F1 => not 100% compatible to OAEI Evaluation

## Draft of Master Thesis Outline

See PDF

## Results for the conference Dataset

Baseline 1:

* Precision: 0.7311210575916458
* Recall: 0.4972549319219703
* F-Measure: 0.5798708616596907

Baseline 2:

* Precision: 0.7174820805515189
* Recall: 0.5526189155525444
* F-Measure: 0.6109677025909451

Outlier Detection Matcher:

* Precision: 0.7706599617313904
* Recall: 0.6316530627157897
* F-Measure: 0.6749185343075025

Details see attached Pivot Table in the excel spreadsheet