MINISTRY OF ECONOMY
DEPARTMENT OF STATISTICS MALAYSIA
NEWSLETTER

DOSM/CTADR/1.2024/SERIES 19

COMPILATION OF TRADE BY ENTERPRISE CHARACTERISTICS (TEC)

What is TEC?

- Integrates two different statistics domains: the International Merchandise Trade database and Malaysia Statistical Business Register (MSBR).
- One of the department's initiatives under the Statistics Big Data Analytics (STASBDA) in embracing Big Data Analytics.
- The Export Import Statistics by State has been produced on a monthly & yearly basis starting December 2019.
- The annual publication can be found on the eStatistics website (https://newss.statistics.gov.my).

TEC Products

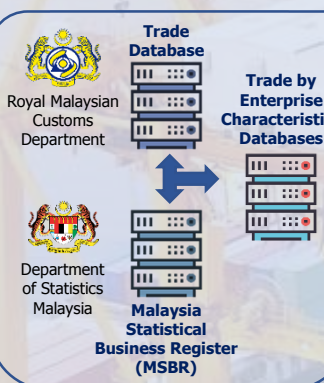
- Monthly basis : Export Import Statistics By State

- Yearly basis : Malaysia External Trade Statistics by State

- Future release : 12 June 2024

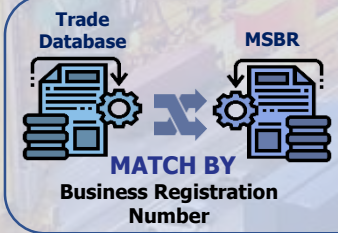
Integration of High Volume of Structured Data with Fuzzy Matching Technique

1.0 DATA INTEGRATION



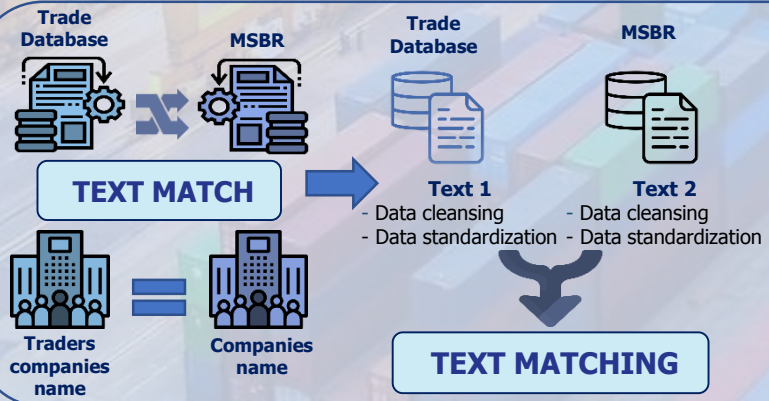
TEC integrates high volume of administrative data from Royal Malaysian Customs Department with Malaysia Statistical Business Register (MSBR). This micro-data linking of MSBR and trade database aims to gain more data insights without initiating a new survey.

2.0 IDEAL SCENARIO

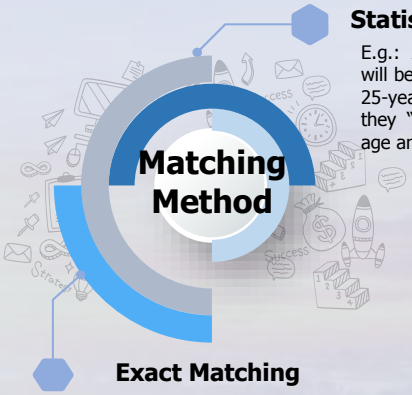


In an ideal scenario, business registration number of exporters and importers can be matched to integrate between MSBR & trade databases.

3.0 TEXT MATCHING



However, due to inaccurate information of the business registration number in the trade database, a text match was performed.
The string matching algorithm was developed to match the name of the traders companies' in the trade database with the MSBR.



Statistical Matching

E.g.: A 25-year-old male will be paired with another 25-year-old male, since they "match" in terms of age and gender.

Exact Matching

E.g.: The information of person A in 1st source will be linked with the information of person A in 2nd source (at record level).

Statistical Matching

Statistical matching is the process of creating a file reflecting the underlying population distribution. Records that are combined do not necessarily correspond to the same entity, such as a person or a business. The files that are being matched can have different units but referring to the same population.

Exact Matching

The goal of exact matching is to link information about a particular record in one file to information on a secondary file in order to create a single file with additional information for each record.

Probabilistic record linkage

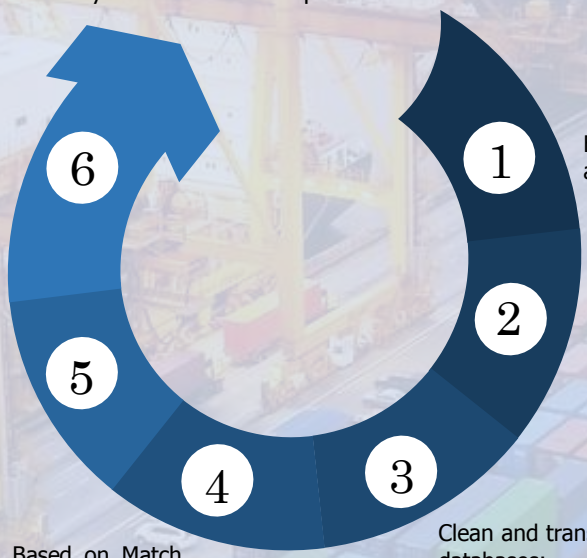
Type of exact matching where there is no unique identifier available for matching. It is done by comparing and quantifying the relative similarity of records in two datasets (it is also known as fuzzy matching). In this method, the similarity score is used.

Deterministic record linkage

ID keys is used to link between 2 data sets. This is the simplest form of record linkage, which produces links based on common identifiers or variables among the available data sources.

Matching Steps

Fuzzy matching is a technique used in data integration and data quality processes to identify and match records that may not have exact matches but are likely to be similar. The process involves several steps, as below:



1. Extract data from trade database and MSBR

2. Determine attributes to match between trade database and MSBR
   - These attributes are the key components that will be used to assess the similarity between records such as Company Name.

3. Clean and transform attributes from both databases:
   - Standardizing formats
   - Removing special characters
   - Converting text to uppercase
   - Handling missing values

4. Test Different Algorithms Based on Match Scope
   - Explore and test various fuzzy matching algorithms. Common algorithms include Levenshtein distance, Jaccard similarity and Soundex.

5. Fine-Tune Algorithms for Refined Matches
   - Adjust thresholds associated with the matching process to optimize results based on the specific characteristics of the data and the desired level of similarity.

6. Create Master/Final Records
   - These records represent the consolidated, cleaned and matched version of the original data.

Performance Evaluation

Threshold Analysis:
Evaluate fuzzy matching performance by analysing the impact of various threshold values on the performance metrics for effective decision-making and improvement.

Conclusion

In conclusion, fuzzy matching is a vital technique for identifying approximate matches in textual or string data, enabling robust data integration and quality processes. It also helps the Department to handle the data processing and provides value added to an existing international trade statistic without initiating a new survey.