Seminar Component

Name of author: Nur Aziha Mansor [1], Muhaimin Naim Md Nasir[2]

Organization: [1] Department of Statistics Malaysia
[2] Telekom Malaysia

Contact address: [1] Department of Statistics Malaysia
Federal Government Administrative Centre
Block C6, Complex C, Lebuh Perdana Selatan
Presint 1, 62514 Putrajaya
Wilayah Persekutuan Putrajaya

[2] Telekom Malaysia
North Wing, Menara TM
Jalan Pantai Baru, 50672, Kuala Lumpur
Wilayah Persekutuan Kuala Lumpur

Contact phone : [1] 603-88857343
[2] 6014-7195535

Email: [1] nuraziha@stats.gov.my
[2] mnaim.mnasir@tm.com.my

Title of Paper

TRADE BY ENTERPRISE CHARACTERISTICS: MANAGING STRUCTURED DATA IN A BIG DATA ENVIRONMENT

Abstract

The amount of data is growing rapidly at a quicker pace in various formats and sources in line with the rise of Big Data. New sources from Big Data have provided an opportunity for the Department of Statistics Malaysia (DOSM) to enhance its efficiency in providing statistical service. As the producer of national official statistics, DOSM has embarked a big data analytic project namely STATSBDA in December 2016. The aim of the project is to get an advantage of Big Data technology and to use alternative sources along with new techniques in producing official statistics. The STATSBDA project is made up of a high volume of structured and unstructured data.

The revolution of Big Data is closely associated with unstructured data. However, there is still a predominant place for structured data that cannot be ignored. The purpose of this paper is to highlight how DOSM managed large structured data in a Big Data environment. Trade by Enterprise Characteristics (TEC) module has been initiated under STATSBDA project. The focus is to integrate between Malaysia Statistical

Business Register (MSBR) with international merchandise trade database in order to add value to the existing trade statistics without conducting new surveys.

In an ideal scenario, business registration number (BRN) of exporters and importers can be matched to integrate between these two databases. However due to poor quality of BRN information in trade database, the similarity between two text strings i.e. exporters and importers companies' name in MSBR and trade database is measured. This paper describes the method used in integrating two databases which includes data cleansing, data standardization and data matching. In short, the technique used in dealing with a high volume structured data in Big Data platform has facilitated DOSM in integrating two different statistical domains.

Key Words: *Big Data, Structured Data, Data Management, Trade by Enterprise Characteristics, Statistical Business Register*

I. Contents

II. Introduction

In this information age, we are overwhelmed with data and the risk of running out of information is very minimal. However, data escalates exceeding the capacity of traditional computing. The growth of data goes beyond relational databases and traditional data warehouse platforms. It needs to be incorporated with technologies that are suited to process, store, analyse, interpret, consume, and transform those data into actionable information.

The increasing focus on collecting and analysing Big Data is shaping new platforms that combine the traditional data warehouse with big data systems in a logical data warehousing architecture. As part of the process, one must decide what data must be kept for compliance reasons, what data can be disposed of and what data should be kept and analysed in order to improve current business processes or provide a business with a competitive advantage. This process requires careful data classification so that ultimately, smaller sets of data can be analysed quickly and productively.

On top of that, Big Data holds tremendous potential on which policy making can be based. It becomes a fundamental to policy making and governance in today's growing information society. There is a need to formulate, evaluate, and implement policies that not only mitigate the risks, but also maximize the benefits of using big data for policy analysis. Thus, accessing to Big Data sources and work with Big Data is becoming important to national statistical systems. The statistical community has recognized the importance and potential use of big data for official statistics. Accordingly, investment needs to be initiated to communicate the advantages of exploiting the wealth of available digital data. New tools and method for capturing, managing and processing are required to take full advantage of Big Data sources.

The Big Data journey of the Department of Statistics Malaysia (DOSM) begins in December 2016. As the producer of national official statistics, DOSM has embarked on a big data analytic project namely Statistics Big Data Analytics (STATSBDA). DOSM realizes that STATSBDA initiative gain much more value out of DOSM’s efforts and it will bring tremendous value to current statistical systems. The STATSBDA project aims to enhance the government's ability to make informed and evidence-based decisions, develop talent in big data analytic as well as responding to the critical needs of the country's transformation agenda. STATSBDA implementation has changed the generic business process from conventional methods in data collection, compilation, analysis and dissemination to a more comprehensive modern method of producing new indicators and insights for official statistical production.

Three main modules which include structured data and unstructured data are undertaken in STATSBDA initiatives. The first module is Trade by Enterprise Characteristic (TEC) which involves structured data. The high volume of administrative data from Royal Malaysian Customs Department was integrated with Malaysia Statistical Business Register (MSBR) which is maintained by DOSM. Micro-data linking of MSBR and trade database was to gain more data insights without initiating new survey. The second module which involves unstructured data; Price Intelligence (PI) was carried out to modernize the price data collection tools with by adopting web scraping techniques to crawl price data from identified websites. Public Maturity Assessment on Official Statistics (PMAOS), the third module which is also an unstructured data project provides DOSM with holistic and comprehensive insight of perception developed by media. This effort gives valuable reference to DOSM to take the necessary action based on publics' perceptions and reactions towards DOSM. Each of STATSBDA modules has its own methods and challenges in managing the data. According to Taylor (2018), structured data is far easier for Big Data programs to digest, while the myriad formats of unstructured data create a greater challenge. Yet both types of data play a key role in effective data analysis.

The aim of this paper is to highlight the method used in TEC module (structured data) which involves data cleansing, data standardization and data matching. The fundamental issue in merging MSBR with trade database is key identifier (ID) i.e. business registration numbers (BRN) are different between dataset. Hence the similarity of two values between two texts strings i.e. exporters and importers companies' name in the databases is measured. The following section of this paper explains briefly about TEC module and Section 4 provides an overview of selected literature that explains on data cleansing, data standardization and data matching. Section 5 provides a detailed treatment of a real data set. Section 6 presents the results and finally, the paper ends with a conclusion.

III. Trade by Enterprise Characteristics in a Big Data Platform

Trade by enterprise characteristics takes a look at international trade statistics from a very specific point of view i.e. the characteristics of the enterprises actively engaged in exporting and importing. Traditional trade statistics record what types of goods are traded across borders between countries but they do not describe the characteristics of the enterprises that are behind these trade flows. In order to know the actor actually engaged in cross border trade, trade data should be linked to the information of enterprises. This identification information can be obtained from SBR, such as name and address, main economic activity of businesses, employment size class, turnover, age of enterprise etc. The linkage of trade statistics with business registers provides description of those who are engaged in global market, and what are their characteristics. In short, the integration of two different statistical domains is an alternative to provide value added to an existing international trade statistics without having to conduct a new survey.

Due to the volume of trade data which is generated at increasing rates, DOSM decides to harness the massive amounts of trade data using big data platform.

The integration of MSBR and trade database is beyond the ability of DOSM current hardware and software tools to process the data within a stipulated time. Thus, this TEC project has become one of the STATSBDA initiatives.

TEC involves combining data from two different sources, which are stored using different technologies. The integration provides a unified view of the data where TEC can provide new information that would not exist in stand-alone statistical domains. The integrated datasets can indicate enterprises that are engaged in international trade as part of global value chains and measure the importance of those firms in the overall economy. Basically, trade database consists of export and import declarations approved by the Royal Malaysian Customs Department and Free Zone Authorities. A company or business may have more than one export and import transactions in trade database.

There are two statistical units used in compilation statistics in Malaysia or maintained in MSBR i.e. establishment and enterprise. The United Nations Statistical Office has defined an establishment as 'ideally, an economic unit which engages, under a single ownership or control, i.e. under a single legal entity, in one, or predominantly one, kind of economic activity at a single location'. While an enterprise is defined as the entire economic activity operating under a single legal entity and it may consist of more than one establishment. Since statistical unit in trade records is enterprise, the integration is done at enterprise level.

The identification of enterprise entities in both MSBR and Customs declaration is business registration number which is maintained by Companies Commission of Malaysia (CCM). CCM is a statutory body that incorporate companies and register businesses as well as regulate companies and businesses in Malaysia. This business registration number has been used as a unique identifier for Malaysia's TEC project since it is very reliable in terms of matching approach.

However, the free text fields in MSBR and Customs declaration forms tend to cause data quality problems particularly business registration number. For that reason, the matching technique is further improved by string matching algorithm approach. Instead of merely using business registration number in matching process, the businesses registered or trading name is also used. Therefore, the element of data management has been considered in the development of STATSBDA architecture which includes data cleansing, data standardization and data matching.

IV. Literature Review

Managing data is the first step towards handling the large volume of both structured and unstructured data. The power of data and the insights gain to make the data useful can only be harnessed through data management best practices. It is important to acknowledge that good data management results in better analytics. By properly managing and preparing the data for analytics, organisations can optimize the Big Data. Managing data is not simple since the data quality and integrity need to be upheld. Inaccurate data can have an impact on results. The quality of the decisions is only as good as the quality of the data used. Thus, data cleansing is a valuable process that can help organizations increase their efficiency. It is a task to ensure data is as accurate and current as possible. Organizations may find that data cleansing enable them to remain compliant with the standards set. Once the data is cleaned, it can be used confidently for deep analysis and for more insights.

Rouse (2010) from WhatIs.com defined data cleansing as a process of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated. An organization in a data-intensive field might use a data cleansing tool to systematically examine data for flaws by using rules, algorithms, and look-up tables. Different approaches will address different issues in data cleansing. According to Maletic and Marcus (2000), general methods for data cleansing include statistical outlier detection, pattern matching, clustering, and data mining techniques. When those methods were implemented on a large data set, the results showed that some of the methods could be successfully applied to real-world data, while others need fine-tuning and improvement. Each of the proposed methods has strength and weakness.

Usually data cleansing is applied when several databases are merged. Merging large databases that are acquired from different sources with heterogeneous representations of information has become an increasingly important and difficult for many organisations especially when identifying lists of names and addresses. The similarity is determined by comparing attribute values with some string similarity and combining the individual similarity values to derive a match decision for a pair of records. Scores of erroneous data sets might happen due to data entry mistakes, faulty sensor readings or more malicious activities.

Hernández and Stolfo (1998) in their study concerned on pre-processing data sets. Large databases were partitioned into clusters such that the potentially matching records are assigned to the same cluster. They used term cluster in line with the common terminology of statistical pattern recognition. Kirsten et al. (2010) proposed two partitioning strategies for generating record match tasks that can be executed in parallel. The first approach was based on the Cartesian product of data set while the second was based on combination of blocking and parallelization. Both partitioning approaches aim at avoiding memory bottlenecks and load imbalances for the resulting match tasks.

The objective of clustering is to figure out commonalities and designs from the large data sets by splitting the data into groups. However, standardization before clustering algorithm leads to obtain a better quality, efficient and accurate cluster result. It is also important to select a specific standardization procedure, according to the nature of the datasets for the analysis. Mohamad and Usman (2013) suggested Z-score as the most powerful method that give more accurate and efficient result compare with decimal scaling and min-max standardization methods.

In short, data pre-processing i.e. data cleansing and data standardising has to be applied to the input databases prior to data matching in order to achieve high quality of matched data. Then the comparison of two or more data sets can be done by emphasising on the various approximate string comparison techniques i.e. data matching. Zhu and Ungar (2000) present a flexible approach to string edit distance, which can be automatically tuned to different data sets and can use synonym dictionaries. Using dynamic programming to calculate string edit distances provides a powerful approach to determining similarity of items described by the strings. This string edit-based matching tool is easily adapted for a variety of different cases when one needs to recognize which text strings from different information sources refer to the same item such as a person, address, medical procedure or product.

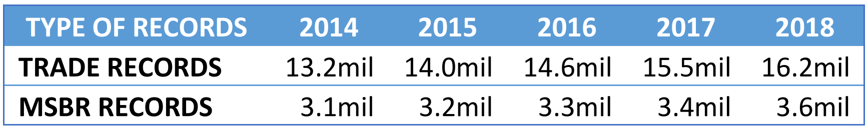
V. Methodology

A. Data Sources

Malaysia trade data was obtained from International Trade Statistics Division which was originally from Royal Malaysian Customs Department. The raw trade data was compiled and processed by International Trade Statistics Division to provide statistics on Malaysia's international trade performance. Then, the processed trade data was handed over to the STATSBDA project team for further processing to generate TEC data.

The TEC data was available once trade data was integrated with MSBR. MSBR is a fundamental property in maintaining the comprehensive list of businesses and companies operating in Malaysia. The integration was able to provide more data insights in order to enrich the international trade statistics by providing closer views of traders.

On average, about 14.7 million of trade records and 3.3 million MSBR records were processed annually started 2014 to 2018. The record shows that the trade data keep on increasing over the years (Figure 1).



TYPE OF RECORDS	2014	2015	2016	2017	2018
TRADE RECORDS	13.2mil	14.0mil	14.6mil	15.5mil	16.2mil
MSBR RECORDS	3.1mil	3.2mil	3.3mil	3.4mil	3.6mil

Figure 1: No. of trade and MSBR records

B. Data Limitation

The TEC data is subject to limitation. Only electronic trade data is going through data management process while the data which is obtained from hard copies are not involved with data cleansing, data standardization and data matching process. This is due to information availability; only selected variables are captured from hardcopies. The information like companies' information and addresses are not captured in the trade system. However, trade values from hard copies are added at the end of the process to ensure the published total trade value is similar with TEC. Based on the record (Figure 2), the trade values from hard copies records are getting smaller as the year increases.

EXPORT: IN TERMS OF TRADE VALUE (IN RM)

	2014	2015	2016	2017	2018
ELECTRONIC RECORDS	96.4%	95.1%	97.8%	99.3%	99.5%
HARD COPIES RECORDS	3.6%	4.9%	2.2%	0.7%	0.5%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

IMPORT: IN TERMS OF TRADE VALUE (IN RM)

	2014	2015	2016	2017	2018
ELECTRONIC RECORDS	97.6%	94.6%	98.0%	99.2%	99.6%
HARD COPIES RECORDS	2.4%	5.4%	2.0%	0.8%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

EXPORT: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
ELECTRONIC RECORDS	99.6%	93.9%	97.3%	99.8%	99.6%
HARD COPIES RECORDS	0.4%	6.1%	2.7%	0.2%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

IMPORT: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
ELECTRONIC RECORDS	95.9%	93.7%	97.9%	99.8%	99.6%
HARD COPIES RECORDS	4.1%	6.3%	2.1%	0.2%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Figure 2: Contribution (%) of trade value by type of records

C. Data Cleansing

In order to achieve goals and meet expectations on how MSBR and trade data can benefit through TEC, data clean-up need to be executed. Data cleansing is undertaken to ensure data is correct, consistent and useable by identifying any errors or corruptions in the data, correcting or deleting them. According to Jones (2017), data cleansing is the crucial step in ETL (extract, transform and load) platform. The objective is to remove invalid characters or phrases that exist in the data. Most of them appear due to the nature of data entry performed by human, which prone to inadvertently entering the undesired data. These invalid characters are mostly neither giving any meaning and nor useful during the matching method. Removing them will not affect the purpose of that data. The examples of unwanted characters present in the exporters' and importers' companies name are:

- "?" (Question mark)
- ";" (Semicolon)
- "$" (Dollar sign)

On top of that, character encoding is also one of the reasons why the data need to be cleaned. Findings showed that some of the raw files were prepared in different encoding that when it is dumped into the database; the data appear differently than it should be. The solution is to ensure the character encodings for both file and database are compatible.

The other scenario is that, there are bad characters appear invisible in the data mimicking whitespaces. It looks normal when the data is viewed from the database but it is otherwise when it is cleaned using the ETL platform. The bad characters are categorised as unprintable characters which are not giving any meaning. According to ASCII, table unprintable characters are other than the decimal value from 32 to 127 (Figure 3). Removing them or replacing them with whitespace will also not affect its purpose.

Character	Hex	Decimal	Character	Hex	Decimal	Character	Hex	Decimal
	20	32	**@**	**40**	64	**`**	**60**	96
!	**21**	33	**A**	**41**	65	**a**	**61**	97
"	**22**	34	**B**	**42**	66	**b**	**62**	98
#	**23**	35	**C**	**43**	67	**c**	**63**	99
$	**24**	36	**D**	**44**	68	**d**	**64**	100
%	**25**	37	**E**	**45**	69	**e**	**65**	101
&	**26**	38	**F**	**46**	70	**f**	**66**	102

'	**27**	39	**G**	**47**	71	**g**	**67**	103
(	**28**	40	**H**	**48**	72	**h**	**68**	104
)	**29**	41	**I**	**49**	73	**i**	**69**	105
*	**2a**	42	**J**	**4a**	74	**j**	**6a**	106
+	**2b**	43	**K**	**4b**	75	**k**	**6b**	107
,	**2c**	44	**L**	**4c**	76	**l**	**6c**	108
-	**2d**	45	**M**	**4d**	77	**m**	**6d**	109
.	**2e**	46	**N**	**4e**	78	**n**	**6e**	110
/	**2f**	47	**O**	**4f**	79	**o**	**6f**	111
0	**30**	48	**P**	**50**	80	**p**	**70**	112
1	**31**	49	**Q**	**51**	81	**q**	**71**	113
2	**32**	50	**R**	**52**	82	**r**	**72**	114
3	**33**	51	**S**	**53**	83	**s**	**73**	115
4	**34**	52	**T**	**54**	84	**t**	**74**	116
5	**35**	53	**U**	**55**	85	**u**	**75**	117
6	**36**	54	**V**	**56**	86	**v**	**76**	118
7	**37**	55	**W**	**57**	87	**w**	**77**	119
8	**38**	56	**X**	**58**	88	**x**	**78**	120
9	**39**	57	**Y**	**59**	89	**y**	**79**	121
:	**3a**	58	**Z**	**5a**	90	**z**	**7a**	122
;	**3b**	59	[	**5b**	91	**{**	**7b**	123
<	**3c**	60	\	**5c**	92	**|**	**7c**	124
=	**3d**	61	**]**	**5d**	93	**}**	**7d**	125
>	**3e**	62	**^**	**5e**	94	**~**	**7e**	126
?	**3f**	63	_	**5f**	95	**Delete**	**7f**	127

Figure 3: ASCI Table

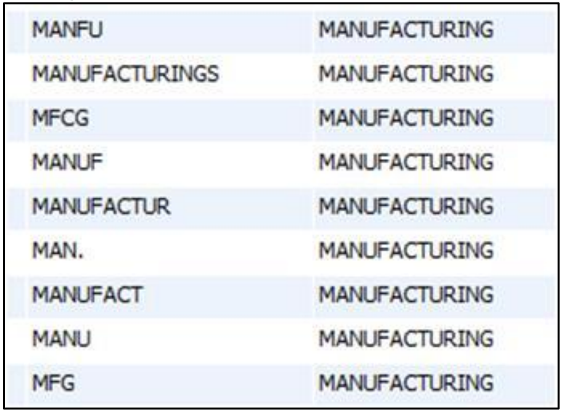
Data profiling is done prior to data cleansing process in order to determine the list of invalid characters and their frequencies. The list is referred as a dictionary where the majorly contains two sets of function namely cleansing and standardization. The latter is to be explained in the next section. JavaScript Regular Expressions is used in searching for pattern that is registered in the dictionary. It is very powerful and highly customizable searching technique to search based on the defined pattern of the invalid characters.

D. Data Standardization

In the context of exporters and importers companies' name, there are many terms that represent a single meaning. According to Jones (2017), profiling is to analyse the data to verify their consistency. In this process, it shows why data profiling is important in data standardization not only in data cleansing. Multiple forms of words exist in the data are mainly due to the using of short forms. For example, the other forms for word:

- “Company” can be “Co”
- “Limited” can be “Ltd”.

Since part of the data entry activities are performed by human, there are tendency for a person to use any form that they are used to. Apart from that, there is also the possibility for the data to be misspelled during data entry is performed. Figure 4 shows the other forms of word "Manufacturing" found during profiling.



MANFU	MANUFACTURING
MANUFACTURINGS	MANUFACTURING
MFCG	MANUFACTURING
MANUF	MANUFACTURING
MANUFACTUR	MANUFACTURING
MAN.	MANUFACTURING
MANUFACT	MANUFACTURING
MANU	MANUFACTURING
MFG	MANUFACTURING

Figure 4: Other forms of word "Manufacturing"

Data standardization objective is to follow one standard term whenever there are multiple forms of word available. This process involves replacing all the non-standardized terms with the standardized term. Searching is also performed by using JavaScript Regular Expressions as some of the term is grouped under the same pattern. On top of that, it is also important to not simply overwrite all the matches as this might change the meaning of the data.

Example for company name *“EMAS CHEMICALS (MAS) CO LTD”*, with the assumption of an entry from the dictionary to replace *“MAS”* with *“MALAYSIA”*

When enforce to only apply for word inside a parenthesis will become:
EMAS CHEMICALS (MALAYSIA) CO LTD

When apply to all occurrences (ignoring the parenthesis) will become:
EMALAYSIA CHEMICALS (MALAYSIA) CO LTD

There are two data standardization categories. The first one is “postfix” which is defined as the words that are usually placed at the end of the company name such as the location name of the company (Example: Melaka, Pahang, Kelantan, Kuala Lumpur). The other one is “generic” which is defined as common name used by companies to name their industries or class of products such as “Retail” and “Transportation”. The aim of having these categories is to prevent the unique name of the company to be modified as some of the unique company name presents in the generic category. The standardization rules are registered in the dictionary to be obeyed by each data that meets the criteria. As a result, each cleaned and standardized data will become as follows:

Before	:	*MAJU LOGISTICS &AMP; FOODFARE S/B*
After	:	*MAJU LOGISTICS & FOODFARE SDN BHD*

E. Data Matching

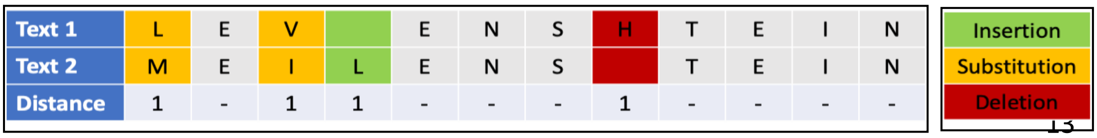
Data cleansing and standardization is really important to simplify the process of data matching. The complexity is drastically reduced by the data cleansing process where there are no unwanted characters left in the data, and the fact that the data has been normalized during the data standardization method. Without these two methods, the matching rate will become lower.

Data matching aims to integrate multiple datasets and in this study is to merge trade data with MSBR. In this process, MSBR is considered as a lookup where exporters and importers companies' name field in trade data will attempt to match with. Elasticsearch is a very powerful service that provides fast searching capability. In addition, Elasticsearch also has it built-in intelligent fuzzy search where the query will sort the best match and most relevant result on top of the list. The first result from the list is initially considered as the closest match possible. However, it does not guarantee that it is accurate as it needs a text matching method to verify the accuracy of matching between the query and its result.

In text matching, it is important to have a good algorithm to compare the two datasets. Levenshtein algorithm is chosen over the algorithms such as Soundex and Hamming algorithm. According to Megter (2016), Soundex is a phonetic algorithm for indexing names by sound as pronounced in English. Thus, Soundex is not fit to use to as the text matching algorithm in this study due to there are company name using the local language which is non-English. Hamming algorithm is less appropriate to use as it is focuses equal length of two texts and also widely used to compare binaries and numbers.

Levenshtein in the other hand is very straight forward text matching algorithm where to texts does not have to be equal in length. Definition by Babar (2018) from Dzone.com, Levenshtein distance is a string metric of calculating the difference of two words. The distance is said as the minimum number of single-character edits (i.e. insertions, deletions, or substitutions) required to change one word into another.

For example, Figure 5 shows the total distance of word "LEVENSHTEIN" to exactly match with the word "MEILENSTEIN" is 3.



Text 1	L	E	V		E	N	S	H	T	E	I	N
Text 2	M	E	I	L	E	N	S		T	E	I	N
Distance	1	-	1	1	-	-	-	1	-	-	-	-

🟩 Insertion
🟨 Substitution
🟥 Deletion

Figure 5: Calculating edit distance between words "LEVENSHTEIN" and "MEILVENSHTEIN"

The distance is then used in automated accuracy checking calculation hence to obtain the matching score. Parameter such as confidence threshold is defined as a benchmark score. In this study, the confidence threshold is set to 88% and the matching rules are as follows:

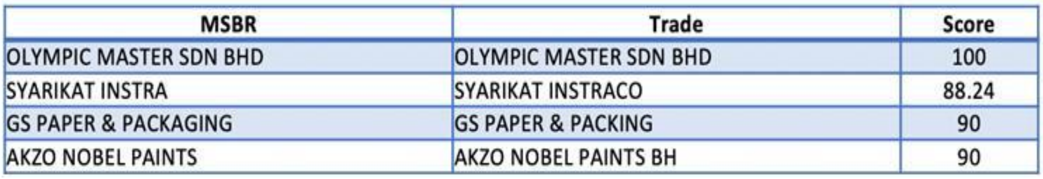
- For score of 100% is considered as exactly match
- For score less than 100% and greater or equal 88% is considered as partly match
- For score less than 88% is considered as not match

In order to handle multiple wordings which mostly do not have an equal length, the accuracy is calculated based on the density of the two texts. Therefore, length of both texts is considered to get the acceptable minimum distance. Any distance between two texts that is less or equal to this minimum distance is considered as similar texts. The acceptable minimum distance is calculated as such:

$$\frac{Maximum\ length\ (text1, text2)\ *\ (100\ -\ confidence\ threshold)}{100}$$

To get the accuracy percentage, the calculation is derived as follows:

$$100 - \frac{levenshtein\ distance\ (text1,\ text2)}{1maximum\ length\ (text1,\ text2)} * 100$$



MSBR	Trade	Score
OLYMPIC MASTER SDN BHD	OLYMPIC MASTER SDN BHD	100
SYARIKAT INSTRA	SYARIKAT INSTRACO	88.24
GS PAPER & PACKAGING	GS PAPER & PACKING	90
AKZO NOBEL PAINTS	AKZO NOBEL PAINTS BH	90

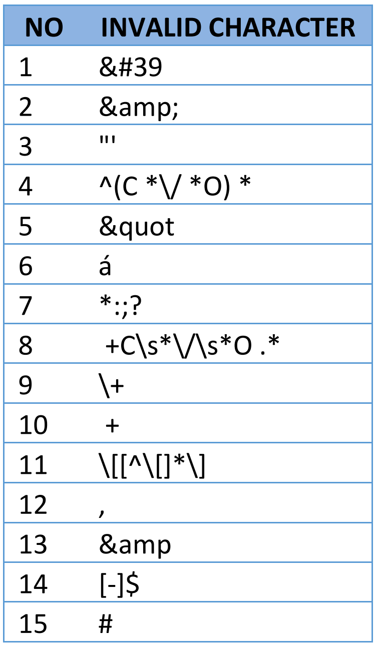
Figure 6: Sample of accuracy scoring percentage

For any score that near to 88%, there are possibilities of false positive results or in other words, these data are supposedly categorized as match. For this case, these data are set to undergo Quality Assurance and Quality Check (QAQC) process where the team need to verify whether to categorized them as match or not. When it is verified as match, it is registered as new rules into the dictionary thus let the platform learn and understand to apply them during the next data cleansing and standardization cycle. In the next matching process, these data will be classified as match.

VI. Results & Discussion

Each of the above-mentioned methods was implemented by using MSBR and trade data from 2014 to 2018. The goal was to prove that these methods can be successfully use to integrate between two different statistical databases. The implementations were designed to work on large data sets where the techniques and algorithms adopted were to reduce the complexity and time taken.

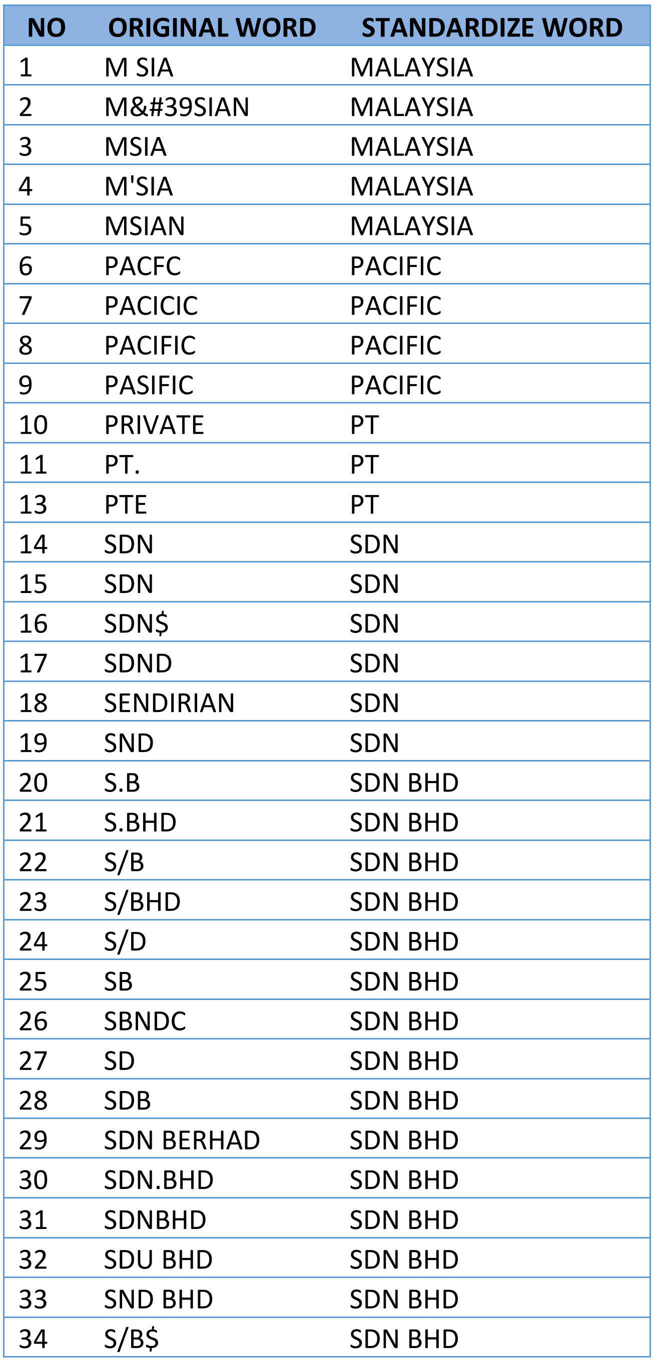
Basically, the techniques for data cleaning may vary from dataset to dataset in order to turn the dataset into a gold mine value. In this study, prior to data cleansing, we first do data profiling to review content and quality in order to prioritize data cleansing and standardization tasks. After getting an overview on the data set content, we then proceed with data cleansing where invalid characters i.e. any character that is not a word character like [^\w\.@-] are removed and replaced with character sets as needed. Figure 7 exhibits the unwanted characters that has been identified in the database and replace with empty string.



NO	INVALID CHARACTER
1	&#39
2	&amp;
3	"'
4	^(C *\/ *O) *
5	&quot
6	á
7	*:;?
8	+C\s*\/\s*O .*
9	\+
10	+
11	\[[^\[]*\]
12	,
13	&amp
14	[-]$
15	#

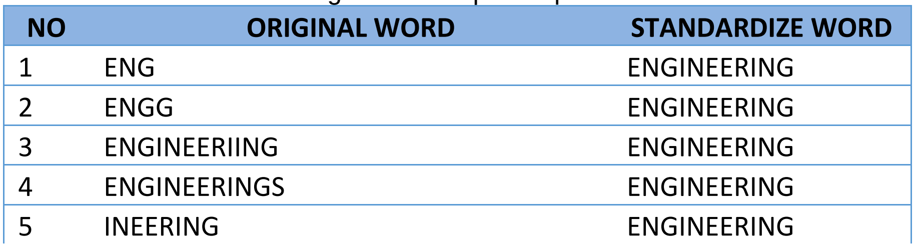
Figure 7: Sample of invalid character

To bring data into a common format, we perform data standardization. In other word, it is a process to correct and harmonize the data i.e. exporters and importers companies' name. This is about to ensure the data is internally consistent. Ideally, data standardization should be performed during data entry but for some reason this is not possible. Thus, a comprehensive back end process is necessary to eliminate any inconsistencies in the data. The standardization process is an important prerequisite when performing data matching. About 100 “postfix” words and 400 “generic” words have been standardized. The examples of those words are in Figure 8 and Figure 9:

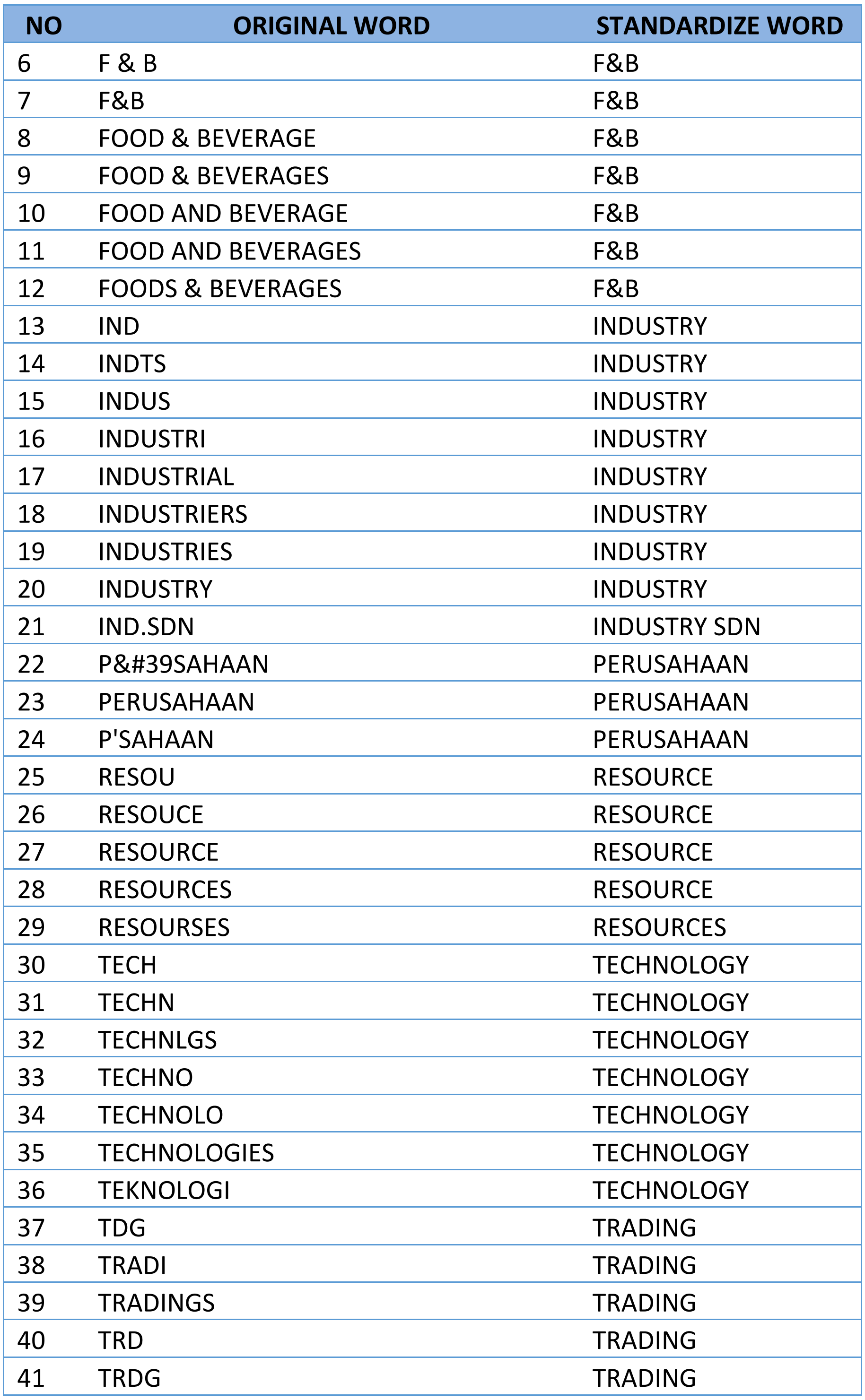


NO	ORIGINAL WORD	STANDARDIZE WORD
1	M SIA	MALAYSIA
2	M&#39SIAN	MALAYSIA
3	MSIA	MALAYSIA
4	M'SIA	MALAYSIA
5	MSIAN	MALAYSIA
6	PACFC	PACIFIC
7	PACICIC	PACIFIC
8	PACIFIC	PACIFIC
9	PASIFIC	PACIFIC
10	PRIVATE	PT
11	PT.	PT
13	PTE	PT
14	SDN	SDN
15	SDN	SDN
16	SDN$	SDN
17	SDND	SDN
18	SENDIRIAN	SDN
19	SND	SDN
20	S.B	SDN BHD
21	S.BHD	SDN BHD
22	S/B	SDN BHD
23	S/BHD	SDN BHD
24	S/D	SDN BHD
25	SB	SDN BHD
26	SBNDC	SDN BHD
27	SD	SDN BHD
28	SDB	SDN BHD
29	SDN BERHAD	SDN BHD
30	SDN.BHD	SDN BHD
31	SDNBHD	SDN BHD
32	SDU BHD	SDN BHD
33	SND BHD	SDN BHD
34	S/B$	SDN BHD

Figure 8: Sample of “postfix” word



NO	ORIGINAL WORD	STANDARDIZE WORD
1	ENG	ENGINEERING
2	ENGG	ENGINEERING
3	ENGINEERIING	ENGINEERING
4	ENGINEERINGS	ENGINEERING
5	INEERING	ENGINEERING

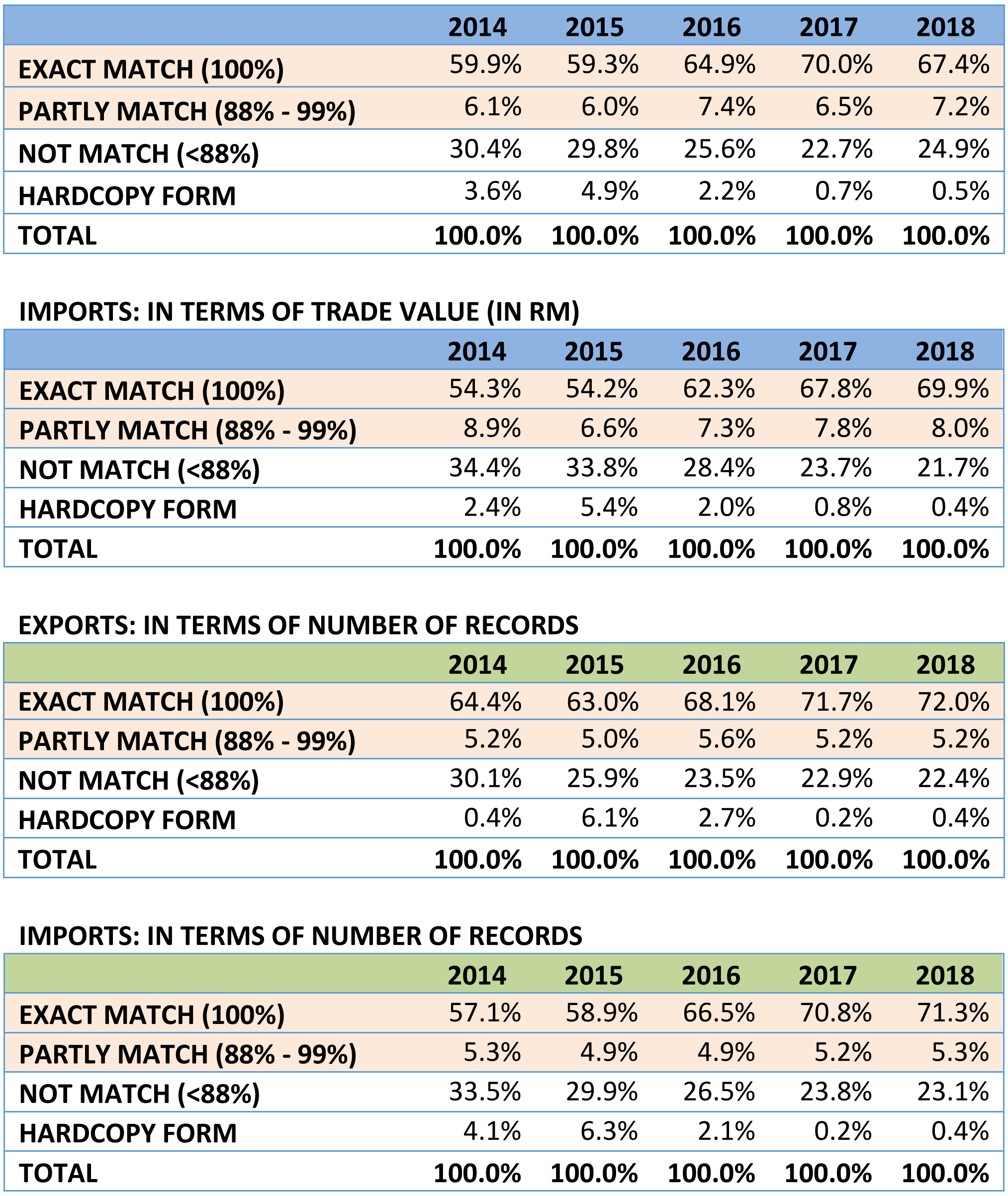


NO	ORIGINAL WORD	STANDARDIZE WORD
6	F & B	F&B
7	F&B	F&B
8	FOOD & BEVERAGE	F&B
9	FOOD & BEVERAGES	F&B
10	FOOD AND BEVERAGE	F&B
11	FOOD AND BEVERAGES	F&B
12	FOODS & BEVERAGES	F&B
13	IND	INDUSTRY
14	INDTS	INDUSTRY
15	INDUS	INDUSTRY
16	INDUSTRI	INDUSTRY
17	INDUSTRIAL	INDUSTRY
18	INDUSTRIERS	INDUSTRY
19	INDUSTRIES	INDUSTRY
20	INDUSTRY	INDUSTRY
21	IND.SDN	INDUSTRY SDN
22	P&#39SAHAAN	PERUSAHAAN
23	PERUSAHAAN	PERUSAHAAN
24	P'SAHAAN	PERUSAHAAN
25	RESOU	RESOURCE
26	RESOUCE	RESOURCE
27	RESOURCE	RESOURCE
28	RESOURCES	RESOURCE
29	RESOURSES	RESOURCES
30	TECH	TECHNOLOGY
31	TECHN	TECHNOLOGY
32	TECHNLGS	TECHNOLOGY
33	TECHNO	TECHNOLOGY
34	TECHNOLO	TECHNOLOGY
35	TECHNOLOGIES	TECHNOLOGY
36	TEKNOLOGI	TECHNOLOGY
37	TDG	TRADING
38	TRADI	TRADING
39	TRADINGS	TRADING
40	TRD	TRADING
41	TRDG	TRADING

Figure 9: Sample of “generic” word

We finally evaluate the result from data matching process by using Levenshtein text matching algorithm. We merge exporters and importers companies’ name in trade database to the same entity in MSBR. The matching rate based on trade value is shown in Figure 10.

EXPORTS: IN TERMS OF TRADE VALUE (IN RM)



	2014	2015	2016	2017	2018
EXACT MATCH (100%)	59.9%	59.3%	64.9%	70.0%	67.4%
PARTLY MATCH (88% - 99%)	6.1%	6.0%	7.4%	6.5%	7.2%
NOT MATCH (<88%)	30.4%	29.8%	25.6%	22.7%	24.9%
HARDCOPY FORM	3.6%	4.9%	2.2%	0.7%	0.5%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

IMPORTS: IN TERMS OF TRADE VALUE (IN RM)

	2014	2015	2016	2017	2018
EXACT MATCH (100%)	54.3%	54.2%	62.3%	67.8%	69.9%
PARTLY MATCH (88% - 99%)	8.9%	6.6%	7.3%	7.8%	8.0%
NOT MATCH (<88%)	34.4%	33.8%	28.4%	23.7%	21.7%
HARDCOPY FORM	2.4%	5.4%	2.0%	0.8%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

EXPORTS: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
EXACT MATCH (100%)	64.4%	63.0%	68.1%	71.7%	72.0%
PARTLY MATCH (88% - 99%)	5.2%	5.0%	5.6%	5.2%	5.2%
NOT MATCH (<88%)	30.1%	25.9%	23.5%	22.9%	22.4%
HARDCOPY FORM	0.4%	6.1%	2.7%	0.2%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

IMPORTS: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
EXACT MATCH (100%)	57.1%	58.9%	66.5%	70.8%	71.3%
PARTLY MATCH (88% - 99%)	5.3%	4.9%	4.9%	5.2%	5.3%
NOT MATCH (<88%)	33.5%	29.9%	26.5%	23.8%	23.1%
HARDCOPY FORM	4.1%	6.3%	2.1%	0.2%	0.4%
TOTAL	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Figure 10: The matching rate by year

Automated system of data cleansing, data standardization and data matching is undoubtedly able speed up the process and is often seen as a replacement for manual process. However, manual process still has a critical role in the quality assurance. Due to that reason, QAQC system has been set up. Using automated and manual system together will lead to a higher data quality (Wyher, 2016). In our study, 88% matching score is set as our confidence threshold and the records can be considered as match. We found there are possibilities of false positive results when the matching scores near to 88%. Therefore, manual checking from human perspective should be done since automated scripts may not pick up the visual issues.

The system will suggest the closest pair of companies' name in both database and the team need to verify manually whether the companies are the same entities or not. The verification process is based on companies' name spellings as well as addresses. Figure 11 shows the example of QAQC done by the team. Once the verification is done, the system will be able to run automatically in the next data standardization process cycle if the same scenario happens.

NO	ORIGINAL COMPANIES' NAME	STANDARDIZE COMPANIES' NAME
1	ADEL ELECTRONIC COMPONENT MNFG SDN	ADEL ELECTRONICS COMPONENT MANUFACTURING SDN BHD
2	AGILENT TECHNOLOGY LDA (MALAYSIA) SDN BHD TAX AGENT	AGILENT TECHNOLOGIES LDA MALAYSIA SDN BHD
3	AIRBUS HELICOPTRS MALAYSIA	AIRBUS HELICOPTERS MALAYSIA SDN BHD
4	BAERLOCHER(MALAYSIA) AND	BAERLOCHER (M) TRADING AND SERVICES SDN BHD
5	BETAMAK ELECTRONIC (MALAYSIA) SDN BHD	BETAMEK ELECTRONICS (M) SDN BHD
6	BUMI MEDIK ARTIFICIAL LIMB SDN BHD	BUMI MEDIK ARTIFICIAL LIMB CENTER SDN BHD
7	CEPCP ELECTRONIC (MALAYSIA) SDN BHD	CEPCO ELECTRONICS (M) SDN BHD
8	CTL PLAST MANUFACTURING	CTLPLAST. MANUFACTURING SDN BHD
9	DAYA OCI ENERGY SDN BHD	DAYA OCI SDN BHD
10	DYNATEC ENGINEERING SDN	DYNATEC ENGINEERING & TRADING SDN BHD
11	EIMTEC MAITENANCE AND SDN BHD	EIMTEC MAINTENANCE & SERVICES SDN BHD
12	FATEAL MACHINERY INDUSTRY SDN BHD	FATAEL MACHINERY INDUSTRIAL SDN BHD
13	HZ GROUTECH	HZ GROUTECH ENTERPRISE
14	JAPTCH INDUSTRY SUPPLY SDN BHD	JAPTECH INDUSTRIAL SUPPLIES SDN BHD
15	TPLINK DISTRIBUTION MALAYSIASDN BHD	TPLINK DISTRIBUTION MALAYSIA SDN BHD

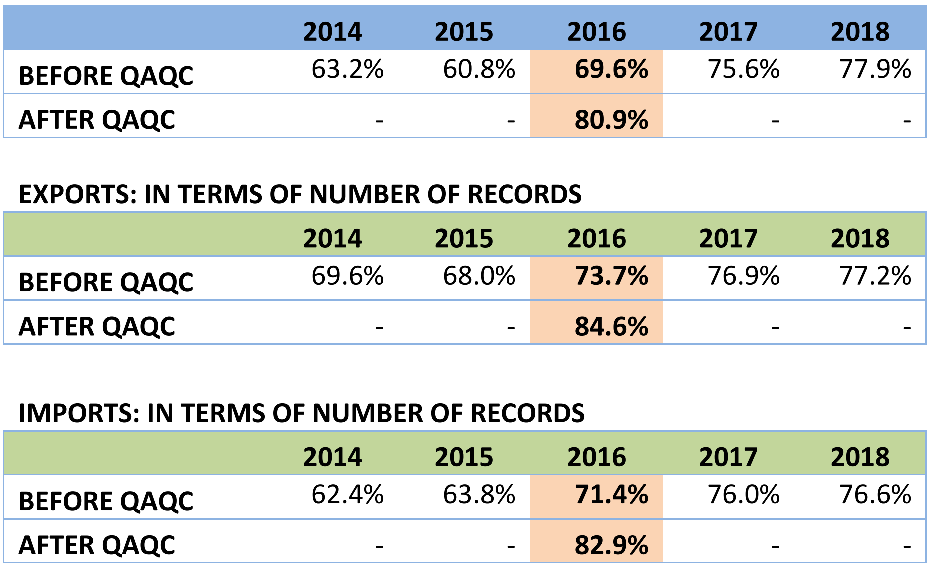
Figure 11: Sample of standardize companies' name

Figure 12 exhibits the matching rate in terms of the trade value before and after QAQC process. As the QAQC process requires human resources, we first put our focus on 2016 data. The result demonstrates that the QAQC procedure able to increase matching rate by 14.8% and 16.4% of exports and imports value respectively. It also proved that the matching rate for 2017 and 2018 increase compared to earlier years since the system is automatically applied 2016 QAQC to 2017 and 2018 data. We believe the matching rate will be higher when the QAQC is carried out on 2017 and 2018 data.

EXPORTS: IN TERMS OF TRADE VALUE (IN RM)

	2014	2015	2016	2017	2018
BEFORE QAQC	66.0%	65.3%	**72.3%**	76.5%	74.6%
AFTER QAQC	-	-	**83.0%**	-	-

IMPORTS: IN TERMS OF TRADE VALUE (IN RM)



	2014	2015	2016	2017	2018
BEFORE QAQC	63.2%	60.8%	**69.6%**	75.6%	77.9%
AFTER QAQC	-	-	**80.9%**	-	-

EXPORTS: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
BEFORE QAQC	69.6%	68.0%	**73.7%**	76.9%	77.2%
AFTER QAQC	-	-	**84.6%**	-	-

IMPORTS: IN TERMS OF NUMBER OF RECORDS

	2014	2015	2016	2017	2018
BEFORE QAQC	62.4%	63.8%	**71.4%**	76.0%	76.6%
AFTER QAQC	-	-	**82.9%**	-	-

Figure 12: Matching rate before and after QAQC

VII. Conclusion

In conclusion, managing structured data in a big data environment is essential for proper data upkeep and to ensure DOSM always produce the most accurate statistics. While data management may seem like a dreaded chore in this project, the platform used makes the task easier and more efficient. Data management is a time-consuming effort, but it can drastically improve the data quality. By managing data properly the first time around, one will be able to confidently rely on the numbers later.

The approach of TEC initiative might be different among National Statistical Offices as it is depending on data availability and quality. With no reliable identifier (ID) to work with, DOSM has decided to venture into text matching algorithm. Starting with data profiling and data cleansing techniques, the exporters and importers companies’ name have been profiled and detected for any invalid character or phrases. Then it is removed or the necessary correction is made. Standardization technique is to fix data irregularities and to get a consistent database so that it can be easily matched with MSBR. Levenshtein distance has been introduced in data matching and the distance is used to obtain the matching score. The record is considered match when the score is 88% and above while any score near 88% will undergo QAQC process. QAQC process involves human judgement in order to decide whether the record is matched or vice versa. Currently, the highest matching rate is about 82% of trade value; the team will explore further in order to improve the matching rate.

VIII. References

[1] Babar, N. (2018, October). The levenshtein algorithm. Retrieved from https://dzone.com/articles/the-levenshtein-algorithm-1

[2] Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media.

[3] Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery, 2(1), 9-37.

[4] Jones, M. T. (2017, December). Working with messy data. Retrieved from https://www.ibm.com/developerworks/library/ba-cleanse-process-visualize-data-set-1/index.html

[5] Kirsten, T., Kolb, L., Hartung, M., Groß, A., Köpcke, H., & Rahm, E. (2010). Data partitioning for parallel entity matching. arXiv preprint arXiv:1006.5309.

[6] Maletic, J. I., & Marcus, A. (2000, October). Data Cleansing: Beyond Integrity Analysis. In Iq (pp. 200-209).

[7] Megter. (2016, January). Fuzzy matching algorithms to help data scientists match similar data. Retrieved from https://www.datasciencecentral.com/profiles/blogs/fuzzy-matching-algorithms-to-help-data-scientists-match-similar

[8] Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303

[9] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[10] Rouse, M. (2010, August). Definition data scrubbing (data cleansing). WhatIs.com. Retrieved from https://searchdatamanagement.techtarget. com/definition/data-scrubbing

[11] Taylor, C. (2018, March), Structured vs. unstructured data. Datamation. Retrieved from https://www.datamation.com/big-data/structured-vs-unstructured-data.html.

[12] United Nation (2007) International Standard Industrial Classification of all Economics Activities (ISIC) Revision 4.

[13] Wyher, T. (2016, October). 5 reasons why manual testing is still very important. Retrieved from https://dzone.com/articles/5-reasons-why-manual-testing-is-still-very-importa

[14] Zhu, J. J., & Ungar, L. H. (2000, August). String edit analysis for merging databases. In KDD workshop on text mining, held at ACM SIGKDD.