Data pre-processing in record linkage to find the same companies from different databases

D Gunawan*, M S Lubis1, D Arisandi1 and B Azzahry1

1Department of Information Technology, Universitas Sumatera Utara, Jl. dr. Mansur No. 9 Kampus USU Medan 20155

*Email: danigunawan@usu.ac.id

Abstract. As public agencies, the Badan Pelayanan Perizinan Terpadu (BPPT) and the Badan Lingkungan Hidup (BLH) of Medan city manage process to obtain a business license from the public. However, each agency might have a different corporate data because of a separate data input process, even though the data may refer to the same company's data. Therefore, it is required to identify and correlate data that refer to the same company which lie in different data sources. This research focuses on data pre-processing such as data cleaning, text pre-processing, indexing and record comparison. In addition, this research implements data matching using support vector machine algorithm. The result of this algorithm will be used to record linkage of data that can be used to identify and connect the company's data based on the degree of similarity of each data. Previous data will be standardized in accordance with the format and structure appropriate to the stage of preprocessing data. After analyzing data pre-processing, we found that both database structures are not designed to support data integration. We decide that the data matching can be done with blocking criteria such as company name and the name of the owner (or applicant). In addition to data pre-processing, the result of data classification with a high level of similarity as many as 90 pairs of records.

1. Introduction

Public services become more complex, especially in licensing, tax and retribution services. All those services cannot be separated and have to be integrated to embody a good governance. The problem arises in Medan city because the company data which have been registered in several public services databases are separated and there is no synchronization among those public services. This may lead to inefficient public services. Therefore, it is required to find the same company the existing records which lie in several databases. In this case, we observed two public services namely Badan Lingkungan Hidup (BLH) and the Badan Pelayanan Perizinan Terpadu (BPPT) of Medan city. This research will discuss the data pre-processing analysis before implementing data matching between two databases.

Previous experiment showed the ability of a data intruder to link records between distinct microdata sets by the co-presence of publicly available aggregate data [1]. This research uses Naïve Bayesian approach as well as used in other research, such as focused crawlers [2][3]. In this research, the Naïve Bayesian approach can avoid the cost for computational process with exact inference. According to this research, Naïve Bayesian algorithm is a useful tool to identify linkages that are most likely to be correct and might be useful for disclosure control.

Another research uses hidden Markov model (HMM) which is trained to standardize typical Australian name and address. These data are collected from health data collection. This research is
intended to solve record linkage on set of simple names [4]. It is required to have comprehensive steps to process longer and complex string. For the address attribute, they found that hidden Markov model produced equal or even better accuracy than rule-based system.

Recent publication proposed secure record linkage for health data integration in developing country [5]. Their technique is claimed can anonymize identifiable private data of the patients. Their experimental results show that the proposed method has successfully linked records in absence of standard identification number in acceptable accuracy for noisy data.

The organization of this paper is as follow: first, the introduction about the background of the research and related works. Next, section two discusses the analysis of data pre-processing to support record linkage. Section three discusses the result of this research. The last but not least is the conclusion.

2. Data Pre-processing Analysis

Data pre-processing analysis was done by conducting several processes. The first process is analyzing the data sources. Both data sources contain many attributes that have different purposes. These attributes should be examined to obtain the similar purpose between them. The similar purpose attributes will be processed later as the parameters to obtain data matching. Second, the data cleaning process for the data in selected attributes from the previous process. Third, we conduct the indexing process or filtering by removing record pairs which impossible to be similar. Fourth, comparing the record pair by using edit distance algorithm. In addition, we conduct data matching by applying support vector machine algorithm. The details of the data pre-processing analysis are as follow:

2.1 Data Source

The data are obtained from two public agencies, which are Badan Lingkungan Hidup (BLH) and Badan Pelayanan Perizinan Terpadu (BPPT) of Medan city. BLH has 500 records which are recorded since 1995 to 2014, meanwhile BPPT has 21,754 records since 2010 to 2014. Each agency has its own database with its own structure. As not all the attributes are related between two databases, we should examine their structure manually to obtain related attributes. Then, the unrelated attributes will be removed and the rest will be kept for the next process.

| Attribute          | Description                                              | Relevance |
|--------------------|----------------------------------------------------------|-----------|
| id_perusahaan      | Unique number for company identification                  | No        |
| nama_perusahaan    | The company name                                         | Yes       |
| nama_pemilik       | The owner or applicant who register the company to BLH    | Yes       |
| alamat_usaha       | The company full address                                  | Yes       |
| nama_kecamatan     | The company district                                      | Yes       |
| nama_kelurahan     | The company sub-district                                  | Yes       |
| telp               | The company official phone number                          | Yes       |
| kodepos            | The company postal code                                   | Yes       |
| no_rekomendasi     | The recommendation number issued by related institution   | No        |
| no_ijin            | The registration number issued by BLH of Medan City        | No        |

The database from BLH has ten attributes as shown in table 1. According to our examination, there are seven attributes that will be used for data matching process such as perusahaan (the company name), nama_pemilik (company owner or person who register the company to BLH), alamat_usaha (the company address), nama_kecamatan (the company district), nama_kelurahan (the company sub-
district), telp (the company phone number), and kodepos (the company postal code). Attributes id_perusahaan (company unique number), no_rekomendasi (the recommendation number) and no_ijin (register number) are not used because of irregular format.

In the other hand, the structure of BPPT database has more attributes. As shown in table 2, the BPPT has 21 attributes. As well as the attributes in BLH database, not all of the attributes in BPPT database can be used. For example, id_perusahaan (company unique number), nofax (attribute to store the company’s fax number), email (attribute to store the company’s email address), jabatan (attribute to store the applicant position), tempatT (attribute to store the applicant residence), tanggal_proses (attribute to store the company registration date), tanggal_ambil (attribute to store the date of license pickup), nama_ambil (attribute to store the person who pick up the license), nomor_identitas (attribute to store the person's identification number), penomoran_ijin (attribute to store license number) are not used because there is no similar attribute (with the same purpose) in BLH database. Attributes which store the province and the city of the company (attribute prop and kota respectively) also removed because all the data only consist of the companies which are located in Medan. According to the difference between both database structures, both databases are not designed to support data integration each other.

| Attribute          | Description                                      | Relevance |
|--------------------|--------------------------------------------------|-----------|
| no                 | Process Id                                       | No        |
| id_perusahaan      | Unique number for company identification         | No        |
| tanggal_proses     | The registration date                            | No        |
| tanggal_ambil      | The date when the license is picked up            | No        |
| nama_ambil         | The person who picked up the license              | No        |
| no_identitas       | The applicant id number                          | No        |
| penomoran_ijin     | The license number                               | No        |
| pemohon            | The applicant name                               | Yes       |
| jabatan            | The applicant position in the company             | No        |
| tempatT            | The applicant residence address                   | No        |
| nama               | The company name                                 | Yes       |
| badan_usaha        | The type of company                              | Yes       |
| alamat             | The company full address                         | Yes       |
| prop               | The company province                             | No        |
| kota               | The company city                                 | No        |
| nama_kecamatan     | The company district                             | Yes       |
| nama_kelurahan     | The company sub-district                         | Yes       |
| kodepos            | The company postal code                          | Yes       |
| notelp             | The company official phone number                 | Yes       |
| nofax              | The company official facsimile number             | No        |
| email              | The company email address                        | No        |

2.2 Data Cleaning

We conduct three steps to clean the data before pre-processing step. Those are handling missing values, smoothing noisy values, and identifying and correcting inconsistent values [6].

3
2.2.1 Handling missing values
There are a few ways to handle missing values. One of them is removing the data which does not have important value in data matching. In this research we removed attribute which is empty or does not have any data. For example, there is no data for attribute kodepos (postal code) in BLH database, meanwhile the kodepos data are exist in BPPT. Therefore, we omitted the attribute kodepos to be included in data matching process. Another way to handle missing value is adding data to the attribute manually. However, the way can only be implemented in numerical data only. This step cannot be implemented in our case because the attribute of the data is string.

2.2.2 Smoothing noisy value
The existing data can contain the false data. For example, the rule for the date format in database is year, month and day. This attribute can consist of the date which has format month, day and year. After examining the data between BLH and BPPT databases manually, we did not found any noisy value that has to be fixed before data pre-processing.

2.2.3 Identifying and correcting inconsistent values
The example of inconsistent values is the sub-district name does not match with the corresponding district. This value should be corrected depending on the availability of supporting data. In several cases, we should prepare the external look-up table or formula/rule that can support data correction.

2.3 Text pre-processing
Data pre-processing refers to process to convert raw data to formatted data which could enhance the data matching result to be more accurate and efficient [4]. As the data is in text format, then this process will include text pre-processing. Dataset is collection of data which depict attributes from real world entities [7]. As this research uses dataset from two databases which have different structure, we should convert them to maintain consistent format. To maintain consistent format in the dataset, we conduct a few steps such as removing unnecessary characters and symbols, normalization, attribute parsing, and verification. These steps are explained as follow:

2.3.1 Removing unnecessary characters and symbols
The first step of data pre-processing is related to the more detailed data cleaning. Values in the dataset might consist of characters, words, terms or abbreviations which do not have any information value. Those values can be removed from the attributes. We also remove unnecessary spaces, such as space in the beginning or the end of attribute values. Characters or symbols are required to be converted to standard form, including convert double spaces into single space. This is required to the normalization and tokenization process. The conversion requires look up table as shown in table 3.

| Table 3. The example of look up table for characters, symbols and abbreviations |
|---------------------|------------------|
| Values              | Conversion       |
| ‘,’ ‘?’ ‘!’ ‘:’ ‘;’ ‘.’ ‘^’ ‘*’ ‘=’ ‘~’ ‘\’ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ ‘‘ '}
2.3.2 Normalization and tokenization

The next step of data preprocessing is normalization and tokenization process. In this step, we convert the value that contains spelling error or typo. For this purpose, we use look up table as shown in table 4. Thus, it is not necessary to consider multiword expressions [8] in normalization and tokenization process.

| Spelling error or typo       | Correct Term                      |
|------------------------------|-----------------------------------|
| 'a r hakim', 'ar hakim', 'a hakim' | arif rahman hakim               |
| 'h m joni', 'hm joni', 'h m jhoni', 'hm jhoni' | moh joni                  |
| 'prof h m yamin', 'prof hm yamin', 'h m yamin', 'hm yamin' | moh yamin             |
| 'sisingamaharaja', 'sisingamaraja', 'sm raja' | sisingamangaraja |
| 'letjend s parman', 'letjen s parman', 'jendral s parman', 'jend s parman', 'sparman' | s parman              |

2.3.3 Attribute parsing

The next step is attribute parsing to form more detailed information. This step extracts an attribute that might contain more than one information. For example, the attribute that store company address information usually consists of more than one information such as street name, alley name, RT/RW (group of several houses within a small area), house/office number, and postal code.

The purpose of this process is obtaining the other information which are available within an attribute value. Parsed attribute value will be used in record pair comparison. Attribute parsing example is shown in table 5.

| Spelling error or typo       | Correct Term                      |
|------------------------------|-----------------------------------|
| 'a r hakim', 'ar hakim', 'a hakim' | arif rahman hakim               |
| 'h m joni', 'hm joni', 'h m jhoni', 'hm jhoni' | moh joni                  |
| 'prof h m yamin', 'prof hm yamin', 'h m yamin', 'hm yamin' | moh yamin             |
| 'sisingamaharaja', 'sisingamaraja', 'sm raja' | sisingamangaraja |
| 'letjend s parman', 'letjen s parman', 'jendral s parman', 'jend s parman', 'sparman' | s parman              |

2.4 Indexing

The purpose of indexing step is to reduce the number of record pairs which will be compared. This step is done by removing the record pairs which are impossible to be matched. Common approach to index the records is by processing all the records and cluster them to several blocks based on certain criteria [9]. These criteria are called blocking key or sorting key. One of the important aspects to index the records does not lie on the method, but the definition of blocking key which cluster the records into similar block [9]. In this research, we used company name, company owner or applicant name, company address and phone number for blocking keys.
2.5 Record Pair Comparison
The next step of data matching process is record pair comparison to calculate the similarity level of both records’ attributes. To compare two attributes, we use Jaro-Wrinkler edit distance. Basically, edit distance is defined as smallest number of insertions, deletions and substitutions which are required to convert a string to another string. The number of editing steps between both strings will be the level of similarity.

2.6 Classification
The last step of data matching process is classification from record pairs candidates as the result of indexing and record pair comparison. The classification is done based on similarity level which are saved in vector format. In this research we use support vector machine to classify records. Generally, the higher similarity level of both records, the higher possibility that both entities are the same.

3. Discussion
As mentioned earlier, the data are obtained from two institutions namely Badan Lingkungan Hidup (BLH) and Badan Pelayanan Perizinan Terpadu (BPPT) of Medan city. The implementation of this research includes data pre-processing, indexing and record comparison.

The data that will be used in data matching are different in structure and content. Thus, these raw data are required to be cleaned and standardized before commencing the next process. The pre-processing data includes removing unnecessary characters and symbols, standardization and tokenization, and attribute parsing.

Indexing process arranges BLH and BPPT data based on blocking key. We used perusahaan (company name) and pemilik (owner or applicant) as blocking key. Each records will be compared to the next step according to similar value in the attributes perusahaan (company) and pemilik (owner or applicant). The record pairs that are not included in indexing are considered as non-matched. After indexing process, all attributes of record pairs will be compared and yield similarity value. After all the data are ready, data classification can be commencing. In this research, we use support vector machine to classify the data to obtain data similarity. As the result, we obtain 90 record pairs generated from 500 BLH records and 21,754 BPPT records.

4. Conclusion
The Badan Lingkungan Hidup (BLH) and Badan Pelayanan Perizinan Terpadu (BPPT) of Medan city are known to share the same companies in both databases. To find the same company, the best way is linking the data with one standard unique key to determine an entity. However, in this case, the unique key to define the same company is not available. Therefore, this research is analyzing the data pre-processing before implementing data matching algorithm. We found that the data are not well prepared for data integration as we found so many differences among both databases. After analyzing data pre-processing we decide to use perusahaan (company name) and pemilik (owner of applicant) as blocking key to support data matching.

References
[1] Smith D and Elliot M 2005 An Experiment in Naïve Bayesian Record Linkage Proc. of the 55th Session of the Int. Statistical Institute
[2] Amalia A, Gunawan D, Najwan A and Meirina F 2016 Focused crawler for the acquisition of health articles 2016 Int. Conf. on Data and Software Engineering (ICoDSE) pp 1–6
[3] Gunawan D, Amalia A and Najwan A 2017 Improving Data Collection on Article Clustering by Using Distributed Focused Crawler J. Comput. Appl. Informatics 1 pp 39–50
[4] Churches T, Christen P, Lim K and Zhu J X 2002 Preparation of name and address data for record linkage using hidden Markov models BMC Med. Inform. Decis. Mak. 2 1 p 9
[5] Khan S I and Hoque A S M L 2017 Health data integration with Secured Record Linkage: A practical solution for Bangladesh and other developing countries 2017 Int. Conf. on
Networking, Systems and Security (NSysS) pp 156–161

[6] Christen P 2012 Data Matching Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection Springer Springer-Verlag Berlin Heidelberg

[7] De Jonge E and Van Der Loo M 2013 An introduction to data cleaning with R Statistics Netherlands

[8] Gunawan D, Amalia A and Charisma I 2016 Automatic extraction of multiword expression candidates for Indonesian language 2016 6th IEEE Int. Conf. on Control System, Computing and Engineering (ICCSCE) pp 304–309

[9] Christen P 2012 A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication IEEE Trans. Knowl. Data Eng. 24 9 pp 1537–1555