26th Meeting of the Wiesbaden Group on Business Registers

- Neuchâtel, 24 – 27 September 2018

Lien Suharni and Rr. Nefriana, Badan Pusat Statistik – Statistics Indonesia

Innovation in Statistical Business Register

The Use of Google Maps Geocoding API and Google Places API Web Service Data for Automation of Updating and Matching Processes in Statistical Business Register

Abstract

After five years in the development of the Statistical Business Register (SBR) in Indonesia, BPS still has the same problem pertaining the limited human resource. The works are mainly done or coordinated by the SBR Secretariat in Sub-Directorate of Statistical Standardization and Classification Development which in charge both for the standardization and classification development also for the SBR development as an addition. While there are only fourteen people in the secretariat, sometimes Subject Matter Areas (SMAs) and interns are placed to help. However, the SMAs have already had some high burden themselves. In 2017, for example, BPS had 122 surveys which mainly conducted by the SMAs (BPS, 2018). This makes it hard for the SBR to be one of the priorities. The interns, on the other hand, will be placed in all around Indonesia territories after only limited times helping SBR. For that constraint, BPS needs some innovations. One of the approaches that can be implemented is the automation. This paper shows two types of research pertaining the SBR automation in updating and matching processes.

For automating the updating process, Google Maps Geocoding API and Google Place API Web Service are used to update and complete the business' data. First, the place identity codes for each sample enterprise are obtained using Google Maps Geocoding API with the name and address of the enterprise as the keywords. Then, based on the place identity codes, the contact data and other data are obtained with both APIs: canonical name, more complete address for the business, latitude, longitude, phone number, active status (active or permanently closed), and website. After that, the number of enterprises that successfully updated with that method is counted and checked whether the update is accurate.

BPS SBR uses many data sources. To avoid duplication, a matching process must be done. Because there is no unique identification number for the business in Indonesia, the matching process is done manually. Actually, a research had been made for automation (Nefriana and Kamaratih, 2017). Despite a large number of units that can be matched automatically on that research, significant numbers of false positive and false negative still existed so that the method has not yet been implemented. Another attempt must be done again because of that reason. In this paper, instead of matching the businesses based on the geocodes, place identity codes are used. In the matching process, if the two businesses have the same business place identity codes based on the response of Google Maps API, then they will be regarded as the same businesses. Besides, the scope of the automation is narrowed down into the enterprises only.

Finally, the number of the enterprises that are successfully matched automatically is counted along with the number of the correct matches and incorrect matches.

Keywords: SBR, SBR automation, SBR matching, SBR updating, SBR maintenance, Big Data, public data, Google Data

INTRODUCTION

In the first five years of the development, BPS SBR has gone through several data integration and updating processes [9,10,11]. The data integration mainly includes uploading data from several sources and then matching them against the data that already stored previously in the SBR statistical unit table [10,11]. Updating process usually also has the same mechanism as the data integration process for batching operation [12]. Furthermore, besides that mechanism, the updating process is also conducted without batching operation [8].

In BPS, the matching procedure (Figure 1) is used to avoid duplication in the SBR statistical unit table [10,11]. Any business entity that will be added to the SBR statistical unit table must be checked first to know if it is already in the SBR statistical unit table. The BPS SBR system finds the top 25 similar business in the SBR statistical unit table to the new incoming business entity. After the operator check the similar businesses one by one –usually including browsing on the internet or calling the business' contact person by phone to get the facts–, the operator will decide if one of the 25 similar businesses is actually the same business with the incoming business. If so, that one similar business will be regarded as the match of the incoming business. The incoming business entity will not be added to the SBR statistical unit table. Instead, the operator has three choices. First, for one or more variables, the operator can edit the old data in the SBR statistical unit table with the new data from the incoming business entity. Second, the operator can replace the old data entirely with the new data from the incoming business entity. Thirdly, the operator can simply leave the old data as they are without using any data from the business entity. The choice depends on the quality of the two sources. Otherwise, if none of the 25 similar business is actually the same entity as the incoming business, the operator will add the incoming data to the SBR statistical unit table as a new entity.

In the case of updating process without batching operation, uploading and matching processes are not done [8]. Instead, the operator will focus on updating data by finding a unit entity he wants and then changing or adding the data of that entity. When he cannot find the unit he wants to edit, he will add it to SBR statistical unit table.

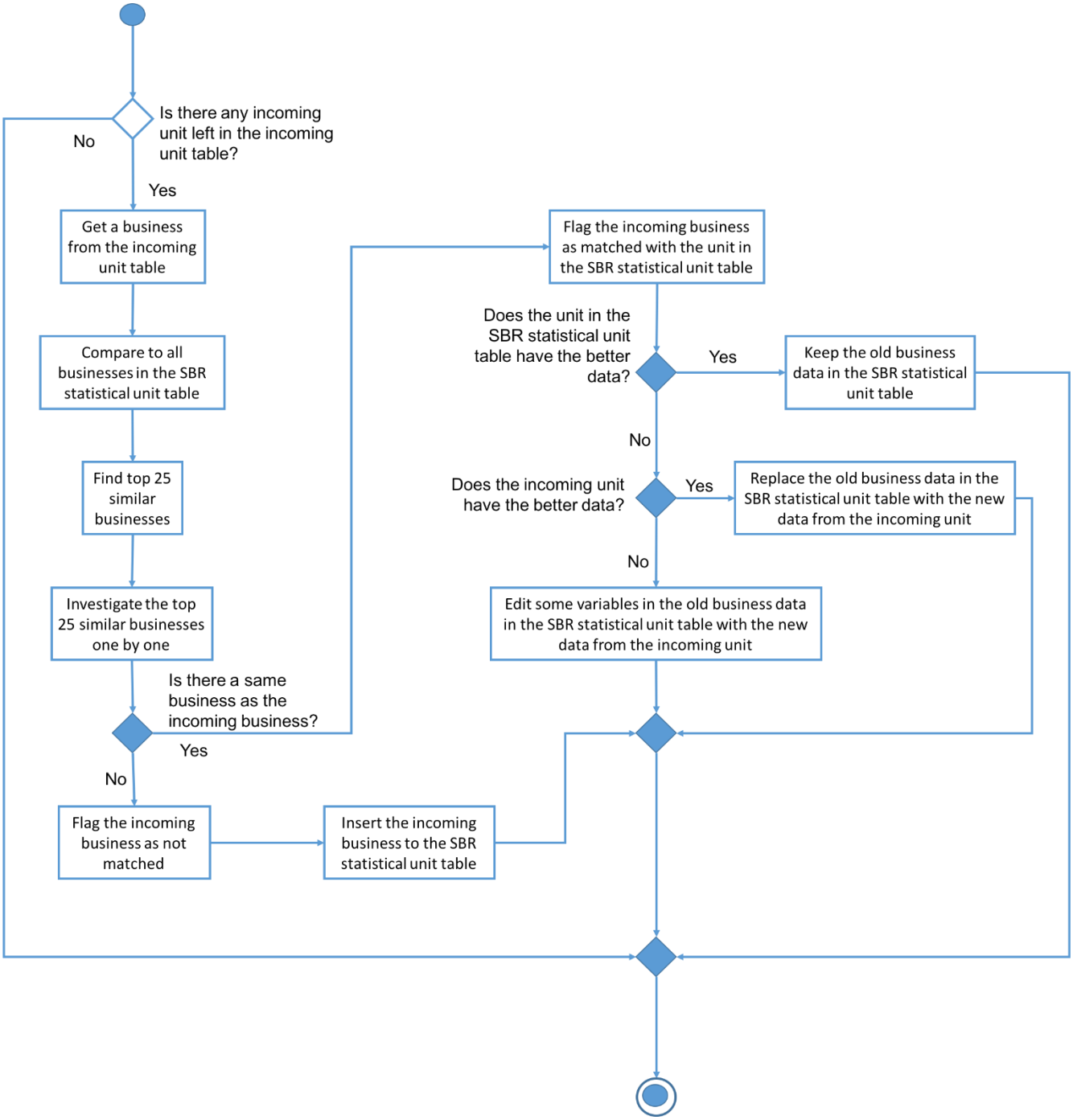

Figure 1. Current SBR Matching Procedure

Data integration and updating are currently very expensive processes for the labor, time, and money required. The works are mainly done or coordinated by the SBR Secretariat in Sub-Directorate of Statistical Standardization and Classification Development which in charge both for the standardization and classification development also for the SBR development as an

addition. While there are only fourteen people in the secretariat, sometimes Subject Matter Areas (SMAs) and interns are placed to help. However, the SMAs have already had some high burden themselves. In 2017, for example, BPS had 122 surveys which mainly conducted by the SMAs [1]. This makes it hard for the SBR to be one of the priorities. The interns, on the other hand, will be placed in all around Indonesia territories after only limited times helping SBR. For that constraint, BPS needs some innovations. One of the approaches that can be implemented is the automation.

LITERATURE REVIEW

A. Empirical Researches

Two kinds of research [10,11] about matching have been done previously. Nefriana, Pahlevi, and Kamaratih (2016) made three efforts to improve the matching facility in the SBR system. They found that tuning the database architecture to be less normal and removing stop words improved the precision/quality of the matching feature and at once improved query time performance. Besides, converting the query from the non-stored procedure to stored procedure decrease the run time from 7.37 seconds to 3.42 seconds which means that the performance was improved. In 2017, Nefriana and Kamaratih made another trial pertaining the matching feature. They used geocodes from Google Maps Geocoding API to do the automation of the matching activity. Despite a large number of units that can be matched automatically on that research, significant numbers of false positive and false negative still existed so that the method has not yet been implemented.

B. The Google APIs

Google has a list of Application Programming Interfaces (APIs) opened for public use [3,13]. They enable some communication mechanisms in their services. Many of the services provide particular data to the public. Parts of the services are Google Maps APIs and Google Places API Web Service under the family of Google Map Platform [13]. One of the products of the Google Maps APIs is the Google Maps Geocoding API [4]. It allows the public to get the geocodes of an address. Beside the geocodes (latitude and longitude), the response of the API includes Google place ID, address components, formatted address, location geometries, and type of the address. Google Places API also gives their user many of the Google Maps Data [5]. The basic response of this API Include (but not limited to) address components, alternative IDs, formatted address, geometry, icon URL, name, the status whether the place has permanently shut down, photo, place ID, the type of the address, URL of the official Google page for the place, number of minutes the place's current time zone is offset for UTC, and vicinity. The contact response includes formatted phone number, international phone number, opening hours, and website. Furthermore, Google Places API provides atmosphere data: price level, rating, and review. In the Google Maps Platform official website [5], it is said that it includes 100 million places and 25 million updates daily.

METHODOLOGY

A. The Steps

To do the research, first (Figure 2) all businesses that were to be matched (incoming businesses) and already detected as enterprises were obtained. There were 2696 businesses detected as enterprises by the profilers previously of all 88324 incoming businesses. Meanwhile, the same way was also done to the businesses that already in the SBR statistical unit table, in which the businesses to be matched (incoming businesses) will search their similar units (potential duplications). In total there were 2364777 businesses.

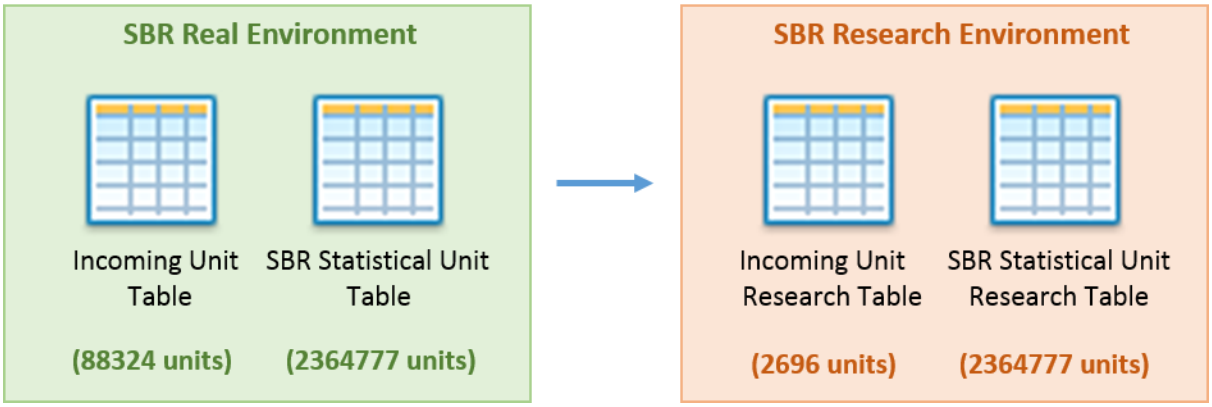

Figure 2. Step 1 of the Research: Getting Research Environment

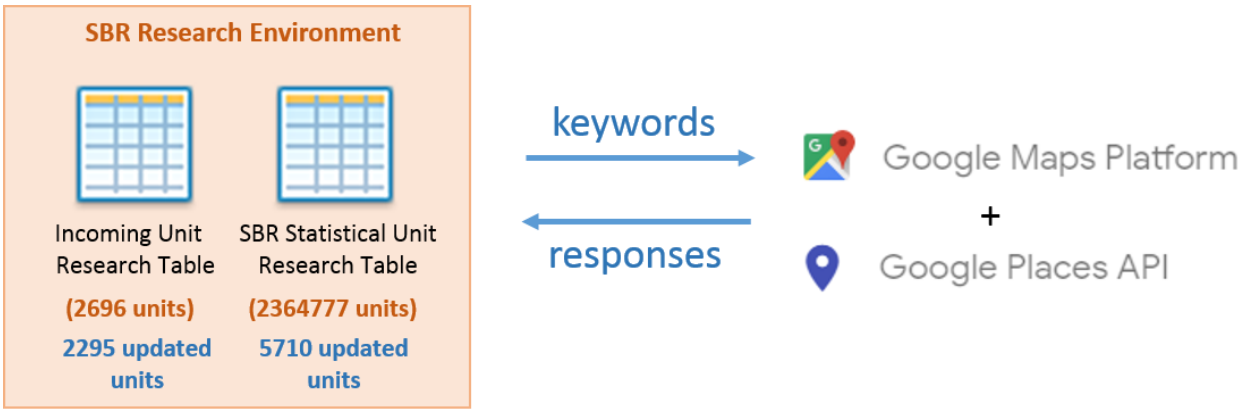

Figure 3. Step 2 of the Research: Getting Google APIs' Data

After that stage, the second stage was getting the data from Google Maps Geocoding API and Google Places API Web Service for both incoming data and the data already in the SBR statistical unit table (Figure 3). The Place IDs for each business were obtained from Google Maps Geocoding API with the keywords containing business name, address, village name, district name, regency name, province name, country name, and its postcode. Then, based on the Place ID, other data were obtained from the Google Places API Web Service. The data are the canonical name, formatted address, latitude, longitude, formatted phone number, website, and the information

whether the business is permanently closed or not. After this attempt, the number of the businesses that the data have been updated was counted for each variable. We got 2295 updated businesses. As for the businesses that already in the SBR data, we got 5710 of 2364777 businesses updated. In this case, the reason was not only because Google cannot give us their API feedback for the rest units, but also because of the time limitation for the research. The number of attempts for getting the APIs' feedback was only 5830 attempts (or 5830 business units) of 2364777. With 101 samples, the number of the businesses that the data have been updated correctly was counted for each variable. By using the internet browser and based on the names and addresses of the entities, the correctness of the updates was checked.

We also had an effort to filter the responses that gave the correct update by only using the API responses which had the same business names (after removing the stop words like "Ltd" "PT" etc. and also all symbols from the business names) as the SBR business names (Figure 4). That way, we only got 24 results from all 101 samples. Again, the number of the businesses that the data have been updated correctly was counted for each variable.

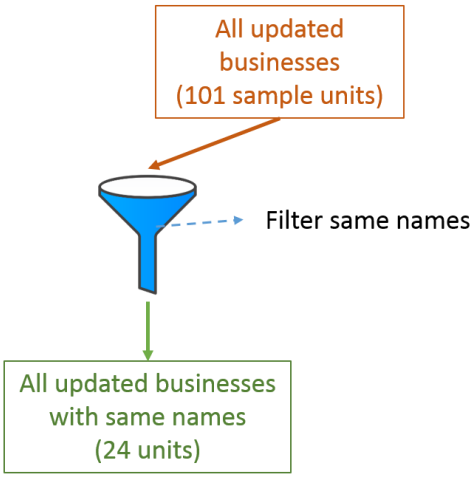

Figure 4. Step 3 of the Research: Filtering Updated Business with Names Similarity Constraint

The fourth stage was running the query trials using SQL Server Management Studio (Figure 5). Here, for each incoming business with their Place ID, we tried to find the corresponding businesses in the SBR statistical unit table with the same Place ID. We also combined the query by considering the similarity between the names of the two businesses, address, phone numbers, facsimiles, and websites.

The kinds of queries that have been tried for auto-matching were (see also Table 1):

A. For each incoming business, find the same business in the SBR table with the same Place ID.
B. For each incoming business, find the same business in the SBR table with the same Place ID and the same industrial categories.

C. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, and the same two digits of International Standard Industrial Classifications (ISIC).
D. For each incoming business, find the same business in the SBR table with the same Place ID and at least have the same phone numbers or the same websites or the same business names or the same addresses or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words.
E. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, and at least have the same phone numbers or the same websites or the same business names or the same addresses or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words.
F. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, the same two digits of ISICs, and at least have the same phone numbers or the same websites or the same business names or the same addresses or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words.
G. For each incoming business, find the same business in the SBR table with the same Place ID and at least have the same phone numbers or the same websites or the same business names or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words. This query was similar to the query in point D, but without considering the similarities between addresses. This was done since it was found that there were the same place identities with the same addresses but actually, they were different businesses. This query tried to compromise that error by considering potential variables that can identify the businesses other than the address information.
H. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, and at least have the same phone numbers or the same websites or the same business names or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words. This query was similar to the query in point E, but without considering the similarities between addresses. This was done since it was found that there were the same place identities with the same addresses but actually, they were different businesses. This query tried to compromise that error by considering potential variables that can identify the businesses other than the address information.
I. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, the same two digits ISICs, and at least have the same phone numbers or the same websites or the same business names or the same facsimile numbers. The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words. This query was similar to the query in point F, but without considering the similarities between

addresses. This was done since it was found that there were the same place identities with the same addresses but actually, they were different businesses. This query tried to compromise that error by considering potential variables that can identify the businesses other than the address information.

J. For each incoming business, find the same business in the SBR table with the same Place ID and at least have the same:
    - both business names and addresses, or
    - both business names and websites, or
    - both business names and telephone numbers, or
    - both addresses and websites, or
    - both addresses and telephone numbers, or
    - both websites and telephone numbers, or
    - both business names and facsimile numbers, or
    - both addresses and facsimile numbers, or
    - both websites and facsimile numbers, or
    - both telephone numbers and facsimile numbers.

    The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words.

K. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories and at least have the same:
    - both business names and addresses, or
    - both business names and websites, or
    - both business names and telephone numbers, or
    - both addresses and websites, or
    - both addresses and telephone numbers, or
    - both websites and telephone numbers, or
    - both business names and facsimile numbers, or
    - both addresses and facsimile numbers, or
    - both websites and facsimile numbers, or
    - both telephone numbers and facsimile numbers.

    The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words.

L. For each incoming business, find the same business in the SBR table with the same Place ID, the same industrial categories, the same two digits ISICs, and at least have the same:
    - both business names and addresses, or
    - both business names and websites, or
    - both business names and telephone numbers, or
    - both addresses and websites, or
    - both addresses and telephone numbers, or
    - both websites and telephone numbers, or
    - both business names and facsimile numbers, or
    - both addresses and facsimile numbers, or

- both websites and facsimile numbers, or
- both telephone numbers and facsimile numbers.

The similarity comparisons between variables are done after removing the symbols, translating them all into the lower case form, and removing the stop words. These stop words were similar to the stop words used in the [7].

Table 1. The Composition of the Trial Queries

	Query											
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Place ID	√	√	√	√	√	√	√	√	√	√	√	√
Business Name				√	√	√	√	√	√			
Address				√	√	√						
Telephone				√	√	√	√	√	√			
Facsimile				√	√	√	√	√	√			
Website				√	√	√	√	√	√			
Business Name & Address										√	√	√
Business Name & Telephone										√	√	√
Business Name & Facsimile										√	√	√
Business Name & Website										√	√	√
Address & Telephone										√	√	√
Address & Facsimile										√	√	√
Address & Website										√	√	√

Telephone & Facsimile										√	√	√
Telephone & Website										√	√	√
Facsimile & Website										√	√	√
Industrial Category		√	√		√	√		√	√		√	√
Two Digits of ISIC			√			√			√			√

The update and the original data must have the same values on this variable

The update and the original data must have the same values on at least one of the variables/combination of variables with this background

We also made the control mechanism for the query with the best result to know whether the place identities contribute to the automation of the matching. That way the control query was the same as the query with the best result, but without considering the place identity similarities.

Figure 5. Step 4 of the Research: Running the Trial Queries

The acceptance of a query was based on the very minimum matching error, preferably zero matching error. After identifying what made the mistakes, we saw the there were actually establishments matched automatically with enterprises and enterprise groups matched with enterprises. With that, it was known that there were many establishments and enterprise groups flagged as the enterprises by the profilers. Then, to know whether filtering the table against the establishments and enterprise groups would improve the results, the incoming unit research table was filtered again against the establishments and enterprise groups (Figure 6 and Figure 7). This was done by removing units that had particular keywords in their name. These keywords were sourced from BPS profiling book that had been purposed to help the profilers in identifying establishments, such as “store”, “plant”, “branch office” and so on. For the enterprise groups,

simply we used keyword “group”. We got 2067 units remained from filtering. After filtering, the number of units in the incoming unit research table was 2067.

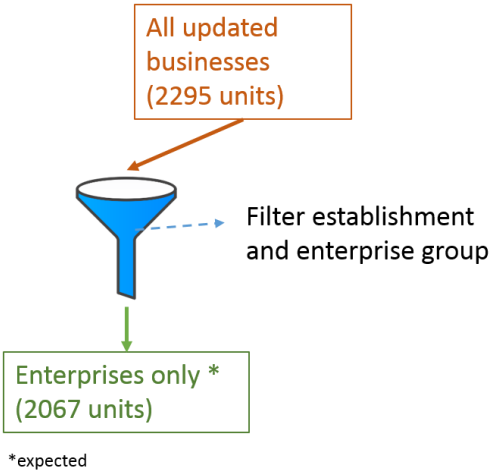

Figure 6. Step 5 of the Research: Filtering Updated Units in Incoming Unit Research Table against Establishment and Enterprise Group Keywords

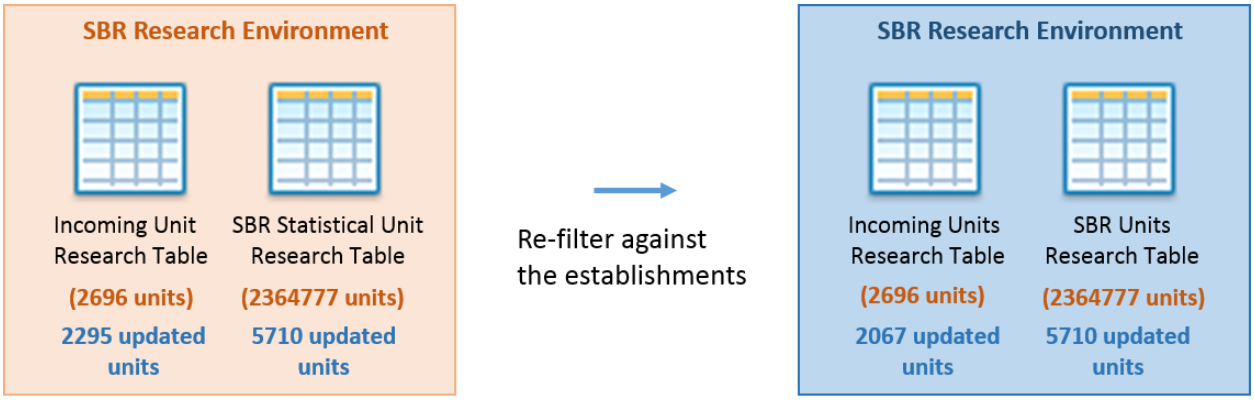

Figure 7. Step 5 of the Research: Filtering Updated Units in Incoming Unit Research Table against Establishment Keywords

The next activity was to run the same previous queries (A to L) on the new incoming units research table again using SQL Server Management Studio (Figure 8).

Figure 8. Step 6 of the Research: Re-running the Trial Queries

After the queries and control query were run, the numbers of correct matches were noted for each queries to know which query was the best.

B. Sampling

To check the proportion of the correct API feedback in the incoming result table without establishment and enterprise group re-filtering, sampling was done because of the research time and resource constraint. We got 101 samples of 2295 business units. Meanwhile, for the matching research, all data was identified one by one without sampling. To know if the update was significant or not, McNemar's test was used with the continuity correction.

RESULT AND DISCUSSION

After the trial, we got 2295 of 2696 incoming businesses' data were updated from the APIs. The reason behind why not all data were updated was that Google cannot give us the API feedback for the 393 units (zero results). As for the businesses that already in the SBR statistical unit table, we got 5710 of 2364777 businesses updated. For this case, the reason was not only because Google cannot give us their API feedback for the rest units, but also because of the time limitation for the research. The number of attempts for getting the APIs' feedback was only 5830 attempts (or 5830 business units) of 2364777.

To asses the result of the API, we got 101 business entities of 2295. Manually we checked whether the updates were correct or not (Table 2) by using the internet browser and based on the names and addresses of the entities (note: for enterprises usually for the SBR we use the head office address as the enterprise address. Except for enterprises that only have one establishment, we use the address of those establishments). For the correct results, the names or addresses between our business data and the update data were actually not always the same. However, we can get the information on the internet when the business' names had changed or the office had moved. We found that 32 (31.68%) of 101 samples got the correct results. If we can differentiate which ones were correct and which ones were incorrect, we can filter the result just for the correct entities. Hence, with that number (31.68%), statistically, the update can be

considered as significant. The incorrect update can be differentiated into two groups: 1) incorrect update, but the correct place; 2) incorrect update and incorrect place. The first group was better than the second one because it gave the information about where the businesses were located correctly so that they were still potential for matching automation.

Later, from 32 correct updates, we also counted the number of update per variable base on 4 classifications (Table 3): the content of the variable was upgraded with the update (upgraded), the content of the variable was downgraded with the update (downgraded), the content of the variable was the same before the update and both had the correct values (same positive), finally the content of the variable was the same before the update and both had the incorrect values or null (same negative). Again, if we can filter the result just for the correct entities, we can say that the update for latitude, longitude, telephone number and website were significant statistically. From the result, we can also see that the API responses successfully filled out entirely the latitude and longitude variables. On the other hand, we got no update at all for the active status. This can be because the business units are actually still alive or Google got no data of the closing units.

Table 2. The Results of Automated Updated Businesses (Sampled)

Type of Update		**Number of Entity**
Correct Update		32
Incorrect Update	Incorrect Update, but the Correct Place	26
	Incorrect Update and Incorrect Place	43

Table 3. The Results of Automated Updated Businesses per Variable (Sampled)

Updated Variable	**Upgraded**	**Downgraded**	**Same-Positive**	**Same-Negative**	**Significant Update? ***
Name	4	1	27	0	Not significant
Address	7	3	22	0	Not significant
Latitude	31	1	0	0	Significant
Longitude	31	1	0	0	Significant
Telephone	8	1	19	4	Significant
Website	11	2	10	9	Significant
Status	0	0	32	0	Not significant

*assumed that we can differentiate between correct and incorrect results

An effort also has been done to filter the update so that only the correct API responses allowed to update the SBR data. Removing the stop words like “Ltd” “PT” etc. and also all symbols from the business names, we then only chosen the API responses which had the same names with our SBR business names. That way, we only got 24 results from all 101 samples (Table 4). However, all of the updates were correct entity updates. Again this number of the update was statistically still significant. However, when we broke down the results to the variable level (Table 5), we only got two variables that the number of the updates were significant: latitude and longitude. Moreover, we also lost the information about the changes to business names in the real world when we used the names similarity filter. We need another research in the future to find the most effective way to know which responses are correct and which ones are incorrect.

Table 4. The Results of Automated Updated Businesses after Filtering (Sampled)

Type of Update		**Number of Entity**
Correct Update		24
Incorrect Update	Incorrect Update, but the Correct Place	0
	Incorrect Update and Incorrect Place	0

Table 5. The Results of Automated Updated Businesses Filtered with Names Similarity Constraint per Variable (Sampled)

Updated Variable	**Upgraded**	**Downgraded**	**Same-Positive**	**Same-Negative**	**Significant Update?**
Name	0	0	24	0	Not significant
Address	5	3	16	0	Not significant
Latitude	23	1	0	0	Significant
Longitude	23	1	0	0	Significant
Telephone	6	1	14	3	Not significant
Website	8	2	6	8	Not significant
Status	0	0	24	0	Not significant

Can be seen in Table 6, we also identified several reasons why particular variables were upgraded or downgraded by the API responses. See Table 5 for the result of the identification.

Table 6. The Identified Reasons for the Changes after Updates by the Google APIs (Sampled)

Variable	Identified Reasons		Notes
	Upgraded	**Downgraded**	
Name	• The real business' name had changed • The old business's name was slightly incorrect		
Address	• The old data had no address information • The address was updated with building number • The address was updated with its street name • The address was updated with its building name and kavling • The old address was incorrect • The old street name was incomplete	• The new kilometer information was incorrect (missing comma) • The address information became less specific (missing building name)	
Latitude	The old data had no latitude information	-	The new latitude information was not rechecked, except for the wrong address updates.
Longitude	The old data had no longitude information	-	The new longitude information was not rechecked, except for the wrong address updates.
Telephone	• The old data had no telephone information • The new telephone number was more specific • The old telephone number was incorrect	• The new website data was missing when the old one had it	

Website	• The old data had no website information • The update had a more specific website information (the old website address was the national level website) • The old website cannot be accessed	• The new website data was missing when the old one has it	
Status			There was no status data updated from the API

For matching automation, we ran all the queries (A-B). From the result in Table 7, we can see that using Place ID resulted in 366 of 2295 can be automatically processed. However, 160 of 336 were actually incorrect matches. Using other variable constraints, i.e. industrial category code and two digits of ISIC, actually helped to reduce the incorrect matches although at the same time reducing the number of the automatic process. Using the business name, address, telephone number, facsimile, and website as some additional constraints also further helped to reduce the incorrect matches while at once reducing the automatic process. Finally, we considered using the place ID with the combination of industrial category code and two digits of ISIC plus having at least two of the former constraints (business name, address, telephone number, facsimile, and website) gave the best results. In this case, query L gave the best result where 110 matching can be automated with only 1 mistake.

Table 7. The Results of Query Trials for Matching Automation

	Query Type											
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Correct Match	260	187	152	244	175	140	237	171	136	180	129	109
Incorrect Match	106	30	15	19	8	3	15	7	2	6	3	1

Although in the last query we only got 1 mistake, it would be better if in the matching process there is no fault. After identifying what made the mistakes, we saw the there were actually establishments matched automatically with enterprises and enterprise groups matched with enterprises. With that, it was known that there were many establishments and enterprise groups flagged as the enterprises by the profilers. Then, to know whether filtering the table against the establishments and enterprise groups would improve the results, the incoming unit research table was filtered again against the establishments and enterprise groups. This was done by removing units that had particular keywords in their name. These keywords were sourced from BPS profiling book that had been purposed to help the profilers in identifying establishments,

such as “store”, “plant”, “branch office” and so on. For the enterprise groups, simply we used keyword “group”. We got 2067 units remained from filtering. We then reran our queries and the result can be seen in Table 8.

Table 8. The Results of Query Trials for Matching Automation with Establishment Filtering

	Query Type											
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Correct Match	224	160	132	212	151	121	206	147	117	153	107	90
Incorrect Match	88	24	9	17	6	2	13	5	1	5	2	0

Table 9 shows that the number of the automatic process was reduced. The good thing was that we found a result with zero error. This was a breakthrough for us after previously trying another research to do matching automation and not finding zero error [10]. If this behavior can stay overtime, this can be the safest way to do the automation. For that, of course, this should be tried again when SBR gets a new data source. The upcoming data from the Indonesia Investment Coordinating Board will give a good chance for that.

To know if Place ID actually helps the automation, we then reran our best query (query L), but this time without considering Place ID. It was found (Table 9) that without Place ID, we got 6 incorrect matches although the number of correct matches increased to 119. With that result, we concluded that query L with Place ID was our best approach for matching automation.

Table 9. The Result of the Best Trial Query for Matching Automation with Place ID Constraint versus Without Place ID Constraint

	Best Query (Query L)	**Control Query (Disregarding Place ID)**
Correct Match	90	119
Incorrect Match	0	6

Currently, the SBR statistical unit table is the product of the integration and updating from five sources: Economic Census 2006 Medium and Large Business data, Subject Matter Areas (SMAs) data, local administrative data gathered by BPS Provincial Offices, profiling data, and Economic Census 2016 data [9,10,11]. In the future, as stated in the BPS SBR Design Document (2017), BPS will also integrate national administrative data in the future [2]. For the maintenance strategy, it will consist of using administrative data, profiling program, survey feedback, profiling program, and Quality Improvement Survey, a survey that will be done to update businesses’ data that have not been updated for awhile by using administrative data, profiling program, or survey feedback. The update using Google Maps APIs is potential in the future for helping the automation of profiling program, which currently done manually by internet browsing or ground check, and also

the automation for Quality Improvement Survey for particular variables. Actually, there are still other variables from the APIs that potentially can be used to update SBR data other than those that have been researched here. However, it still needs more effort in processing and mapping the data before being used to update the SBR data. It can be the next research agenda. Hopefully, with the growth of Google Maps data, the number of data to update will also increase and be better.

Lastly, the matching automation will be potential for integrating SBR data with the Indonesia Investment Coordinating Board in the near future, Tax Office data, Big Data, and other sources of data. This automation can also be done in the non-batch operation; operators can use this matching automation to check whether a particular unit has already been in the SBR statistical unit table before adding one unit to the table.

CONCLUSION

A research has been done on the automation of updating and matching SBR data. Hopefully, the problem of the limited human resource can be helped with these kinds of automation in the future.

The update using Google Maps Geocoding API and Google Places API Web Service data was significantly effective. From 7 variables, two variables were statistically extremely significant for the update: latitude and longitude. If in the future, we can differentiate perfectly which API responses are right and which responses are wrong, telephone and website updates are also statistically significant besides the latitude and longitude updates. We also found that using the place ID with combination of industrial category code and two digits of ISIC plus having at least two of five constraints (business name, address, telephone number, facsimile, and website) gave the best results for matching automation with zero incorrect matches and 4.35% of the data can be matched automatically. That was effective for incoming data that the statistical units have been detected as enterprises.

The update using Google Maps APIs is potential in the future for helping the automation of profiling program, which currently done manually by internet browsing or ground check, and also the automation for Quality Improvement Survey for particular variables. On the other hand, the matching automation will be potential for integrating SBR data with Indonesia Investment Coordinating Board in the near future, Tax Office data, Big Data, and other sources of data. This automation can also be done in the non-batch operation; operators can use this matching automation to check whether a particular unit has already been in the SBR statistical unit table before adding one unit to the table.

REFERENCES

[1] Badan Pusat Statistik. (2018). *Laporan Kinerja Badan Pusat Statistik 2017*. Jakarta: Badan Pusat Statistik.

[2] Colledge, M., Suharni, L., Pertiwi, R. P., Parulian, Y., Utama, G. A., …. Marsinta, A. (2017). STATCAP-CERDAS: Statistical Business Register and Large Business Unit Design Document. Jakarta: Badan Pusat Statistik.

[3] Gerardus Blokdyk. (2018). *Google APIs a Clear and Concise Reference.* Australia: Emereo Pty Limited.

[4] Google. (2018). *Geocoding API: Developer Guide*. Retrieved from https://developers.google.com/maps/documentation/geocoding/intro.

[5] Google. (2018). *Places API: Place Details.* Retrieved from https://developers.google.com/places/web-service/details.

[6] Google. *Google Maps Platform: Places.* Retrieved from https://cloud.google.com/maps-platform/places/.

[7] Kurniawan, N. B., Haryanto, D., Fitriyani, I. A., Widodo, H. M., Adhi, A. A. P., Pertiwi, R. P., … Rahmawati, M. (2016). Statistical Business Register: Modul Pelatihan (Capacity Building) Model Statistik Ekonomi. Jakarta: Badan Pusat Statistik.

[8] Kurniawan, N. B., Haryanto, D., Fitriyani, I. A., Widodo, H. M., Adhi, A. A. P., Pertiwi, R. P., … Rahmawati, M. (2015). Integrated Business Register: Modul 3 Penyusunan Profil Perusahaan (Profiling) 2015. Jakarta: Badan Pusat Statistik.

[9] Nefriana, Rr. (2017). *Laporan Integrasi Data SE2016-Listing dengan SBR*. Jakarta: Badan Pusat Statistik.

[10] Nefriana, R., & Arsiani, I. K. (2017). *The Use of Google Maps API for Optimizing the Matching Process in SBR System.* Bangkok: Asia-Pacific Economic Week 2017.

[11] Nefriana, Rr., Pahlevi, S. M., & Arsiani, I. K. (2016). *Tuning Statistical Business Register System Matching Feature*. Tokyo: 25th Meeting of the Wiesbaden Group on Business Register.

[12] Nuwibowo, A., Arianto, Sedyo, L., Barudin, Kurniawan, N. B., Nefriana, Rr., … Isnawati, I. (2015). *Usaha/Perusahaan Menengah Besar Preprinted 2015*. Jakarta: Badan Pusat Statistik.

[13] Wikipedia. Google APIs. Retrieved from