# Geographic Information System for Higher Education via Data Scrapping

*Agung* Alfiansyah[1,*], and *Anas* Azhar[2]

[1]School of Applied STEM, Prasetiya Mulya University, BSD City Kavling Edutown I.1, Jl. BSD Raya Utama, BSD City, Kabupaten Tangerang 15339
[2]Dinas Pendidikan Kota Palembang, JL. Pramuka No. 929 Km. 5,5 Kel. Srijaya Kec. Alang-alang Lebar, Kota Palembang, Sumatera Selatan 30151

**Abstract**. This paper aims to build a database that can be used as the most important part of a geographic information system by using data from reliable sources. Data collection is done through process scrapping on some official website of Indonesia government. In addition to collecting quantitative data of higher education, this paper also mapping geographically from the high level of existing in Indonesia. It is expected that this data will be able to help the community and government in observing and analyzing the quantity and distribution of universities. The result of this research is one database that is ready to be integrated in more complex geographic information system. In addition, the scrapping method used allows us to perform the upgrade of data that has been obtained previously.

**Key words**: Data scrapping, geographic information system, higher education in Indonesia, spatial data.

## 1 Introduction

Indonesia is a large country with a population of almost $235 \times 10^6$, covering an area of 1 910 931 $km^2$. The country comprises more than 17 504 islands, makes it the largest archipelago in the world. The population is still dominated by young generation, whereby 44.72 % of its population is younger than 25 yr. This is particularly important due to the increasing needs to provide education and employment for the young.

As an emerging economy, Indonesia is considered as a low middle-income country entering the third stage of economic development, called the "efficiency driven economy" by the World Economic Forum (WEF 2012). Indonesia needs to address many complex issues to improve its competitiveness as it makes the transition to a new phase of economic development. In 2014 Indonesia is ranked at 34$^{th}$ by the World Competitiveness Index. Well-educated human resources, excellence in scientific research and better linkages to industry and government are regarded as key policy priorities in nearly all countries in this stage.

To achieve that, higher education programs are offered by five types of institution namely: academy, polytechnic, college, institute and university. The first two are specializing in vocational education stream, whilst the last three are more comprehensive and allowed to offer all education streams. A college (*Sekolah Tinggi*) is a specialized institution focusing on one particular academic discipline. Unlike universities, institutes are specialized in a particular group of disciplines such as sciences and technologies, arts or agriculture.

Universities in Indonesia are largely offered by the private sector. Out of around 3 500 institutions, only around 150 institutions are public (established and operated by the government). The public institutions in Indonesia are mostly under the Ministry of Education and Culture (98 institutions) and Ministry of Religion Affair (52 institutions). In the last 5 yr, the government has also established a number of new public institutions by converting the status of existing private institutions. The database of higher education directory are actually accessible online via https://forlap.ristekdikti.go.id/ [7].

*Corresponding author: agung.alfiansyah@prasetiyamulya.ac.id

## 2 Problem formulation

This paper aims to explain the ways used to obtain a database used to build a GIS that can be used to map the situation of higher education in Indonesia. The data is collected through the scrapping process of the official website of Ministry of Research, Technology and Higher Education (*Kementrian Riset dan Pendidikan Tinggi*) so that all important data can be obtained. In addition to quantitative data, the process offered will also be able to obtain spatial data from the geographical position of the university recorded on the website of the Ministry of Research, Technology and Higher Education.

Ultimately, this study aims to help people gain real insights, values, and benefits quickly, and precisely because much of the information can be explored and then processed and presented better in the form of GIS in its database pages. In addition, this study also aims to find out how to obtain record information/college data in Indonesia that has good and credible accuracy, design efficient data structures to create GIS, and know visualization of geographic data of Indonesian universities to be easily accessible and neatly presented to the general public.

## 3 Related works

This study explains how Wikipedia users choose places by identifying the correlation between conditions and locations based on social media objects Flickr and Twitter. This study focuses on the potential of geographic information on Wikipedia that should be able to customize its documents with the coordinates of photos on Flickr and corresponding Twitter messages, and also allows modeling of experimental languages on Flickr and Twitter that can fit in Wikipedia articles. The result is that the language can be used as a gazetteer-based data retrieval method (used by Yahoo!) and approaches on Wikipedia itself [1].

This study aims to improve the techniques for crawlers through the World Wide Web on search engines that focus on hierarchical crawling techniques, where crawlers are created dynamically at runtimes for different domains with resource sharing essences [2]. Swastika [3] developed a web based GIS for tourism in Gianyar Regency (Bali) that integrated google manually gathered data based to the google maps. For our case, this approach is not applicable as we have a big amount of highly dynamic data.

### 3.1 Geographic's information systems (GIS)

There several definition of GIS definitions and somehow the definition of GIS depends on who uses and processes it, from the user's background and point of view. In short, GIS is a computer system capable of storing data and using it to describe the places on the surface of the earth using the data from Geographic Positioning System [5]. In general, the definition of GIS includes three main components of hardware, software, and spatial data. In general, GIS should able to increase the performance of following tasks:

i). Speed and easiness to access a large amount and volumes of data;
ii). Ability to select data with specific areas or themes;
iii). Ability to analyze the characteristics of spatial data;
iv). Ability to look for certain characteristics in an area and then update data quickly.
v). Modeling alternative data and calculations;

As GIS is designed to store spatial data there are the following type of Information:

i). Latitude and Longitude as reference geography;
ii). Details of relationships as appropriate;
iii). Non-spatial attributes, such as data corresponding to usage.

The functions that exist in the GIS, one of which is there is enter data, storage, management, transformation, analysis and output. In general, outcomes of a GIS are: maps, graphs, addresses, lists and statistical summaries.

### 3.2 Google maps

Google visualization tools used by developers with HTML5 and JavaScript knowledge can already utilize up to 90 % of the overall functionality (Application Programming Interface). Some methods for visualization do not even require full code writing skills. A developer also needs the capability to change the source database in web server including SQL language communication and jQuery.

For data bound by geographic location, map visualization is often considered attractive to use. Unlike the graphite bar and the line drawing, it will be difficult to redraw the geographic area as desired. In visualization, Google Maps provides a basis for spatial data, and there are two types of graphics, namely geo-chart visualization and geo-map visualization. The two types of graphs are a delivery model according to the intensity and quantity of geographic access of each required area.

The difference between geo-map and geo-chart is the method used to render the actual image on the map. The rendering technique that geo-map uses Adobe Flash quickly but lacks the options for certain configurations. While geo-chart is rendered in VML (Vector Markup Language) or SVG (Scalable Vector Graphics), the configuration is detailed [5, 6].

## 4 Proposed method

To obtain credible data, we retrieve the necessary data through the official website pages issued by the technology research and higher education ministries by means of website scrapping. We gather the needed data from two official sites which are: FORLAP Website (Higher Education Databases): https://forlap.dikti.go.id/ [7] and BAN-PT (*Badan Akreditasi Nasional Peguruan Tinggi* — Board of Higher Education Accreditation): https://banpt.or.id/direktori/prodi/pencarian_prodi [8]. The workflow to build databases for higher education application using data scrapping can be illustrated using this scheme:

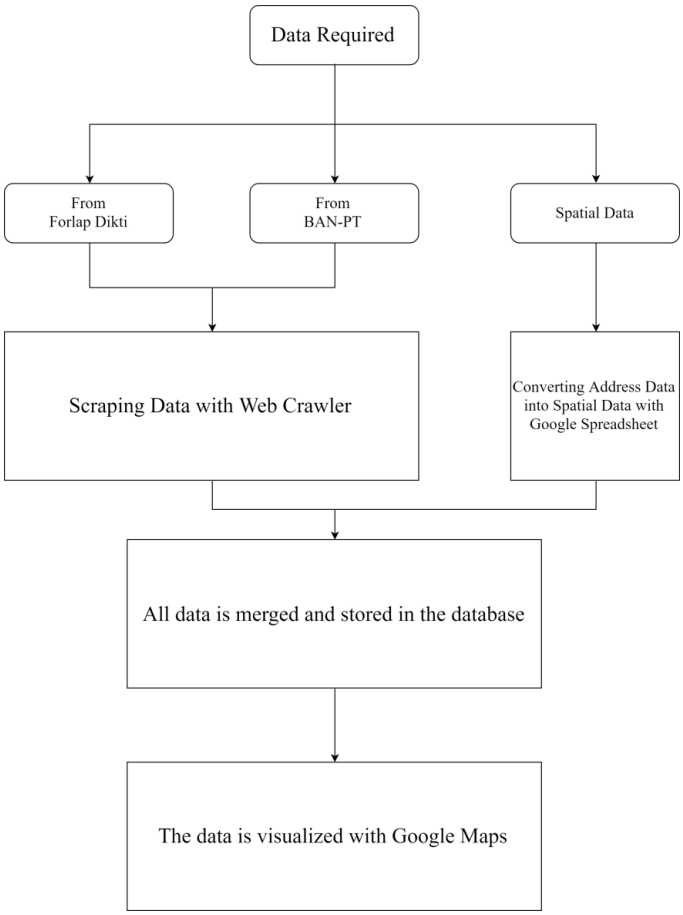


**Fig. 1.** Workflow in GIS database acquisition via website scrapping.

The workflow to build databases for higher education application using data scrapping can be illustrated using this scheme:

**Table 1.** Different data acquired from official governmental websites.

| Data Source | Gathered Data |
|---|---|
| [7] | Institution, University Code, University, Study Program Code, forlap code University Status, Level, Permanent Lecturer, Lecturer with Doctoral Degree, Lecturer with Master Degree, Number of Students, Establishment year, Establishment Letter Number, Date of Establishment, Address, City, Postal Code, Telephone Number, Facsimile, email address, websites, study program url, University url. |
| [8] | Study Program name, accreditation level, Letter reference number, expired date, accreditation status. |

To perform this data retrieval task, we use data retrieval using web crawling. This method is a stage that developed and used by a search engine to download web pages on the internet. Once downloaded, the next process is parsing on a web page and converted into a database on the server. Web crawlers also provide indexing to get valuable information about the web page.

### 4.1 FORLAP data acquisition

In retrieving data process in the FORLAP site, the website page it could not be scrapped only in one process as there are many site map sharing to get various data in the website. Therefore we did the workflow sub-division of the process described in Figure 2. This following chart describe the data collection diagram.


**Fig. 2.** Sub-process to perform data scrapping in FORLAP website.

### 4.2 FORLAP data acquisition

On the BAN-PT website, the accreditation data is displayed by the POST method in the PHP script which is then forwarded in html. This results in the absence of a query set in the site address section of the page. Therefore, we saved the pages of BAN-PT site into the local server in order to be processed and cleaned data. This method should be enhanced for the next application in order to increase the performance of the methods.

### 4.3 Spatial data conversion

The data that has been obtained by scraping and then stored in the database is processed as such, then there is some data that is used to obtain spatial data. This process is intended to convert the data from a university/institute physical address to its geospatial location. This data location is usually given in terms of the value of the position in latitude and latitude of the position on earth. To perform this process, we use the function available in google sheet, ie geocode. So that the address data that has been obtained from the forlap needs to first be tabulated in the spreadsheet to then be transformed automatically.

This spatial data helps us to visualize our gathered data on the online map. Most of conversion result give the right location, but some of them are not exactly in the correct location. This error is caused by the fact that the function cannot find the location with the defined name demanded by user, which is then done in wider location search. However, this type of error can be overcome by doing several experiments with different queries with the same meaning.

## 5 Implementation

To perform the appropriate tasks that have been described in the specifications, we use the following implementations of the following flows:

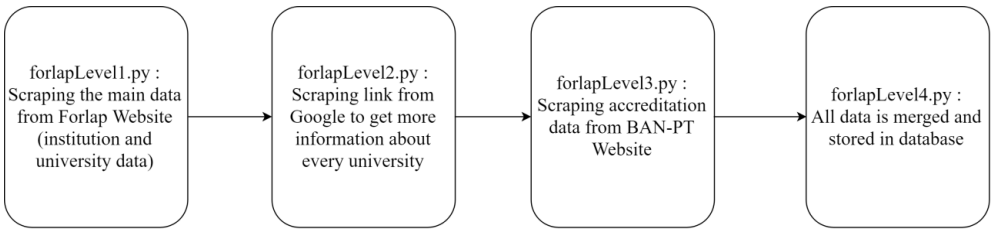


**Fig. 2.** Different level of scrapping in higher education data acquisition.

Overall step task performed by data scrapping in highest level (university), can be describe as follows:

```
Time initialization
Connect to MySQL databse
Access forlap web using urllib
Scrap data using beautifulSoup
Build University dataframe
for each link in dataframe do
  for all data with <tr> tag do
    for all data with in tag <a> href=true do
        get university store to dict.
        get link and store to dict
Build data dict using Pandas
Build data dict in MySQLPandas
Print runtime execution
```

**Fig. 3.** Data scrapping in python-like *psudocode*.

Algorithm used in other parts is not too different from the first part, the difference is there because the structure of web pages are not the same, so it takes a different way of digging the data, although it is actually the same. Next step of the scrapping method is data address conversion from physical address to the latitude-longitude which performed using google sheet web application. Some of the results of this address conversion can be observed in the following figure:

| R | S | T |
|---|---|---|
| Addr | Lat | Long |
| Universitas Gadjah Mada Kec. Depok - Kab. Sleman - Prop. D.I. Yogyakarta 55281 | -7.7713847 | 110.3774998 |
| Universitas Indonesia Kota Jakarta Pusat - Prop. D.K.I. Jakarta - Indonesia 10430 | -6.1949311 | 106.848885 |
| Universitas Sumatera Utara Kota Medan - Prop. Sumatera Utara - Indonesia 20155 | 3.5653554 | 98.6567748 |
| Universitas Airlangga Kec. Mulyorejo - Kota Surabaya - Prop. Jawa Timur 60115 | -7.2698208 | 112.784576 |
| Universitas Hasanuddin Kota Makassar - Prop. Sulawesi Selatan - Indonesia 90245 | -5.13241 | 119.488 |
| Universitas Andalas Kota Padang - Prop. Sumatera Barat - Indonesia 25163 | -0.914559 | 100.4595468 |
| Universitas Padjadjaran Kota Bandung - Prop. Jawa Barat - Indonesia 40132 | -6.893941 | 107.617288 |
| Universitas Diponegoro Kec. Tembalang - Kota Semarang - Prop. Jawa Tengah 50275 | -7.0520829 | 110.4399777 |
| Universitas Sriwijaya Kota Palembang - Prop. Sumatera Selatan - Indonesia 30662 | -2.9846917 | 104.7335316 |
| Universitas Lambung Mangkurat Kota Banjarmasin - Prop. Kalimantan Selatan - Indonesia 70123 | -3.2975189 | 114.5864305 |

**Fig. 5.** Address to geo-location conversion result.

It is rather difficult to check the thoroughness of this conversion result thoroughly. At this stage we only check a few (just a few) of the several universities that we know and examine as far as we know. This method may be very unreliable and inaccurate, but so far if the results obtained from the conversion process still exist in one sub-district area, it is in our opinion still quite good.

## 6 Results

We tested our proposed approach to acquire the data and we have following results (as 17 July 2017):

i). 4 027 universities and higher education institutions across Indonesia with a 2.5 MB SQL data load;
ii). 25 087 study programs from all universities and educational institutions in Indonesia with a load of 15.5 MB SQL data;
iii). 1 130 accredited information of the university and the institute of education of Indonesia with the contents of SQL data 192 KB;
iv). 19 102 Accreditation information from all Study Programs of all Universities and Institutions of Higher Education in Indonesia with a 3.5 MB SQL data load.

That result represents 88.3 % of the scrapped data from the total data in FORLAP website. The rest there is some of the data that cannot be scrapped as they were hidden from google crawler thus, cannot be found using google. It is possible that this happens because that the university has just submitted data into the FORLAP, so it is not indexed in the google search engine. In a short time, the data will be retrieved once google has done the indexing.

And finally, the visualization was performed on the Google Maps so that it can be publically accessed online. Intuitively, this technique makes easier for us to process data on the map. We do not need to build from scratch again or even do not need any API access code on google maps. Google maps themselves already provide layers, so users will be expected to query by checking on the slides tab to get the desired information, although obviously the functionality is limited and cannot be customized anymore. Google itself only gives 2 000 lines for each layer, this makes it difficult for us to add exploratory data on the map, because the data will be inserted more than 20 000 rows. Finally, we only add data that can be inserted in 2 000 line only.

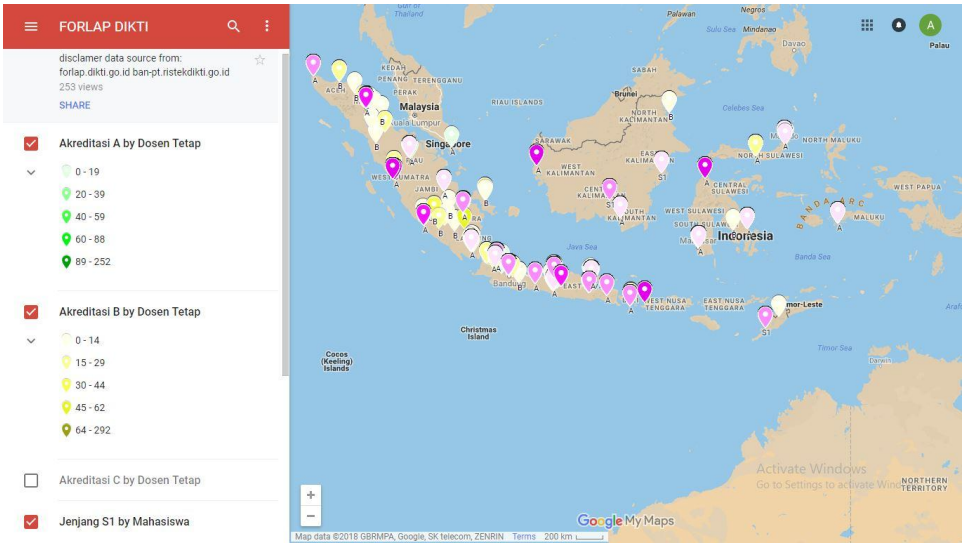


**Fig. 5.** Higher education visualization in google map with simple query.

We consider this limitation is a drawback for that should be solved for the next step of development, we proposed to build a dedicated server to accommodate the amount of data and also capability to perform complex data query and visualization.

## 7 Conclusion

The results of data retrieval using robot crawlers are data obtained from the forlap site pages and BAN-PT as well as data conversion in google docs. Acquisition of data achieved from the total data reached 4 027 of the 4 557 universities registered in the forlap dikti or in percentage reached 88.3 %. Approximately 11.7 % more data cannot be scrapped, this is because data cannot be found when searching using google. Search in google is done to pass through the security gap in the forlap site that is a captcha. Google does not or has not indexed a small portion of the Forlap site page that the crawler requested, so no data was found.

## Annex

Partial visualized data in google maps with limited query can be found in: http://bit.ly/forlapdikti [9]. Source code and acquired data base can be forked in github from https://github.com/azharnian/forlapDikti [10].

## Reference

[1] O.V. Laere, S. Schockaert, V. Tanasescu, B. Dhoedt, C.B. Jones. ACM Transaction on Information System. **33**,2(2014). https://dl.acm.org/citation.cfm?id=2629685
[2] A. Kundu, R. Dutta, R. Dattagupta, D. Mukhopadhyay. International Journal of Intelligent Information and Database Systems, **3**,1:90–106(2009). https://dl.acm.org/citation.cfm?id=1810808
[3] I.W. Swastika. *Sistem informasi geografis berbasis web untuk pemetaan pariwisata Kabupaten Gianyar: Studi kasus pada Dinas Pariwisata Kabupaten Gianyar*. [Web based geographic information system for tourism mapping of Gianyar Regency: A case study at Gianyar Regency Tourism Office]. [Undergraduate Thesis] Teknik Informatika, Universitas Pembangunan Nasional Veteran Yogyakarta (2011). [in Bahasa Indonesia]. 56 http://repository.upnyk.ac.id/905/1/SKRIPSI.pdf
[4] I. Heywood, S. Cornelius, S. Carver. *An introduction to geographical information system*. Third Edition. London: Pearson (2006). 18-27 https://www.amazon.com/Introduction-Geographical-Information-Systems-3rd/dp/B0041W4PSG
[5] T.L. Ruthkoski. *Google visualization API essentials*. Birmingham: Packt (2013). 7-18 https://www.safaribooksonline.com/library/view/google-visualization-api/9781849694360/
[6] B. Fry. *Visualizing data*. Sebastopol: O'Reily Media (2008). 6-14 https://www.amazon.com/Visualizing-Data-Explaining-Processing-Environment/dp/0596514557
[7] Kementrian Riset, Teknologi dan Pendidikan Tinggi, Pangkalan Data Pendidikan Tinggi [Online] from https://forlap.ristekdikti.go.id/ (2018). [Accessed on March 15th 2018]. [in Bahasa Indonesia]

[8] BAN-PT, Direktori Hasil Akreditasi Program Studi, [Online] from https://banpt.or.id/direktori/prodi/pencarian_prodi (2018). [Accessed on March 15th 2018]. [in Bahasa Indonesia]
[9] A. Alfiansyah, A. Azhar. *Visual Forlap Dikti*. [Online] http://bit.ly/forlapdikti [Accessed on March 15th 2018].
[10] A. Alfiansyah, A. Azhar. *Project of scraping web http://forlap.dikti.go.id/ with Python2.7* [Online] from https://github.com/azharnian/forlapDikti [Accessed on March 15th 2018].