ISSN 2085-4579

Implementasi *Web Scrapping* pada Website Eduesia.Com untuk Pengukur Kesenjangan Jumlah Mahasiswa Perguruan Tinggi di Indonesia

Yuri Pramana, Wira Munggana
Program Studi Sistem Informasi, Universitas Multimedia Nusantara, Tangerang, Indonesia
yeo.yuri@gmail.com, wira@umn.ac.id

Diterima 14 Desember 2015
Disetujui 28 Desember 2015

***Abstract*— Education is an important asset for the future of each individual, community, nation, and world. However, not everyone can have the same opportunity as the field of economic and geographic constraints that trigger gaps in the public assessment of the quality of one's intellectual. Based on data provided by UNESCO, Indonesia recorded no. 4 from other ASEAN countries, in the level of *Gross Enrolment Ratio*. The research process is intended to form a picture of educational inequality at the college level in every region in Indonesia. This study was made with Web scrapping's techniques and produce a website that presents a valid and accurate information and analysis of the ratio of *Gross Enrolment Ratio* to show the level of participation at the college level.**

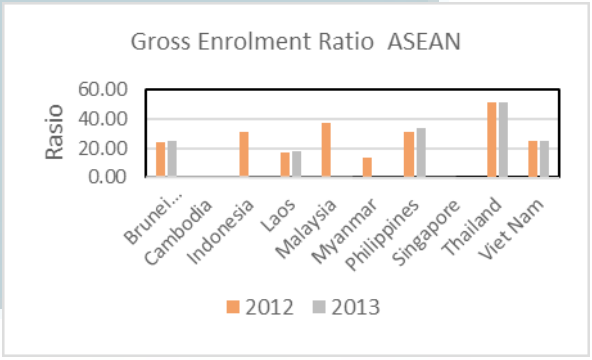
***Index Terms* — Data Collection, UMN, Web Crawling, Data Cleansing, Web Scrapping**

I. PENDAHULUAN

Pendidikan merupakan aset penting bagi masa depan masing-masing individu, masyarakat, negara, dan dunia. Untuk sebuah negara pendidikan menjadi tolak ukur kecerdasan warganya. Sangat penting bagi warga negara untuk mendapatkan hak mengakses dan memiliki tingkat pendidikan yang sama. Berdasarkan perhitungan *Gross Enrolment Ratio* yang dilakukan oleh UNESCO, Indonesia berada di peringkat 4 dari negara ASEAN, seperti dapat dilihat pada Gambar 1.

Penelitian ini dilakukan, untuk mengetahui tingkat partisipasi untuk memasuki perguruan tinggi pada daerah-daerah di Indonesia. Proses pengumpulan data menggunakan *web scrapping, web crawling,* dan *data cleansing* dengan menggunakan data yang disediakan situs pemerintah yaitu BAN-PT, data.go.id, dan dikti.

Dari penelitian ini akan menghasilkan sebuah website dan yang dapat menampilkan informasi kesenjangan tingkat edukasi pada perguruan tinggi yang ada di Indonesia yang diharapkan dapat membantu pemerintah dalam meratakan distribusi pendidikan.



Gambar 1. **Gross Enrolment Ratio ASEAN**

II. LANDASAN TEORI

Data Collection merupakan proses dalam pengumpulan data dan variable, dengan cara sistematis yang berguna dalam menjawab kebutuhan dari perushaan dan organisasi. Dalam penggambilan data pada penelitian ini menggunakan teknik *web scrapping, web crawling,* dan *data cleansing*

Web crawling digunakan untuk mendapatkan kode HTML dari *website* tersebut. Kode HTML merupakan hasil akhir dari proses *web crawling* yang akan dianalisa untuk mencari URL yang terdapat dapat *website*.

Web scrapping merupakan proses yang

melibatkan sebuah dokumen *semi-structured* dari internet, umumnya pada halaman *web* menggunakan *markup language* seperti HTML atau XHTML

Data Cleansing ditujukan untuk pengidentifikasi data yang di dapat untuk memastikan bahwa data relevan dan tidak terdapat kesalahan pada data untuk menghindari adanya kesalahan pada proses analisa data.

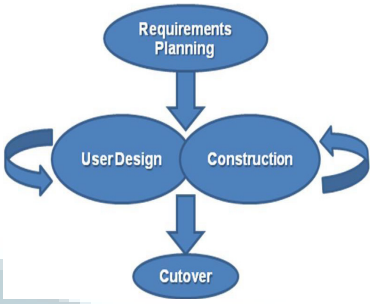
III. METODOLOGI PENELITIAN

Pendidikan tak kalah penting dari kesehatan finansial. Seiring dengan perkembangan zaman, muncul permasalahan baru yang membuat dunia pendidikan semakin dibutuhkan dan penting untuk didapatkan, karena itulah pendidikan semakin diprioritaskan untuk masyarakat. Berdasarkan perhitungan *Gross Enrolment Rastio* yang dilakukan oleh UNESCO, Indonesia berada di peringkat 4 dari negara ASEAN. Meskipun pendidikan Indonesia belum dapat dijajarkan dengan pendidikan di Asia lainnya, namun pendidikan di Indonesia mengalami perubahan yang lebih baik dari sebelumnya.

Perubahan pendidikan di Indonesia dapat dilihat adanya program kebijakan pemerintah wajib belajar 9 tahun, Ujian Akhir Nasional (UAN) yang diwajibkan untuk anak SMA. Atau SMK dan dengan adanya Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) dengan tujuan untuk membangun pendidikan di Indonesia

A. Metode Penelitian

Metode yang digunakan dalam penelitian ini untuk menyelesaikan akar permasalahan yang dihadapi pada penelitian ini adalah penggunaan rancangan aplikasi dengan model *System Development Life Cycle* (SDLC) *Rapid Application Development* (RAD).



Gambar 2. Siklus Metode **Rapid Application Development**

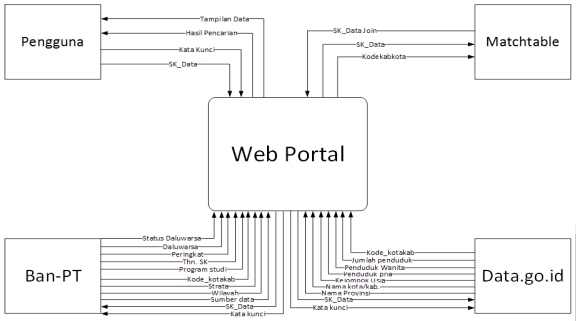
Gambar 2 menunjukkan empat fase dalam RAD yakni *Requirements & Planning, User Design, construction* dan *Cutover*. Berikut merupakan penjelasan dari masing-masing fase.

B. Fase *Requirements Planning*

Requirements Planing merupakan gabungan dari fase *planning* dan fase *analysis* dari metode SDLC. Fase ini bertujuan untuk mengumpulkan *requirement* yang akan digunakan untuk membuat *website*. Pengambilan data dilakukan dengan cara *web scrapping, web crawling,* dan *data cleansing*. *Web Crawling* adalah sistem yang digunakan untuk melakukan pengunduhan pada halaman *website* dalam jumlah besar. *Web Scrapping* merupakan proses pengekstrakan data pada halaman website yang berupa kumpuluan kode HTML, javascript, css, dll. *Data Cleansing* adalah proses pengidentifikasi data yang didapat untuk memastikan data relevan dan tidak terdapat kesalahan, untuk menghindari kesalahan pada proses analisa data.

C. Fase *User Design*

Fase *User Design* juga dikenal sebagai desain tahap fungsional, tahap ini model sistem data dan menggambarkan proses sistem yang terjadi pada *website*. Sistem pada *website* dirancang menggunakan DFD yang sesuai dengan proses dan aliran data dengan 2 tingkat yaitu diagram konteks dan diagram level 1. Gambar 3 mengilustrasikan rancangan DFD yang telah dibuat.



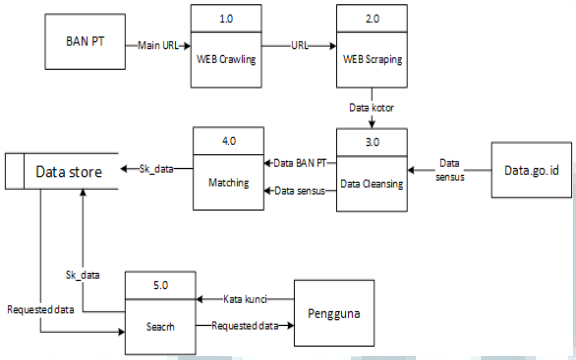
Gambar 3. Diagram Konteks

Keterangan :

Hasil pencarian kepada pengguna merupakan data produk secara lengkap.

Tahap dalam diagram :

- Pengguna ingin melihat informasi terkait dengan data yang cari pada website
- Pengguna mencari data dengan fitur *searching* dengan memilih dan memasukan kata kunci.
- Pengguna ingin melihat data dalam suatu kumpulan sesuai dengan tempat yang mereka pilih
- Pengguna dapat melihat hasil pada web dari hasil yang pengguna cari.



Gambar 4. Diagram Level-1 *Website*

Keterangan:

Data yang terdapat pada Data.go.id merupakan data bersih.

Tahapan pada Diagram Level 1 :

- Proses 1.0 Web crawling. Merupakan proses pengambilan URL pada BAN-PT untuk mengambil halaman pada website
- Proses 2.0 Web Scrapping. Digunakan untuk mengambil data yang di perlukan untuk di analisan dan masih menjadi data kotor.
- Proses 3.0 Data Cleansing. Proses ini kedua data dibersihkan untuk *noise* pada kedua data.
- Proses 4.0 Matching. Pada 4.0 kedua data disamakan entitasnya sehingga data mendapatkan data dari kedua sumber
- Proses 5.0 Search. Proses ini mengambil kata kunci yang diberikan kepada pengguna kemudian diteruskan ke Data Store. Kemudian Data Store memberikan data yang dinginkan dan ke pengguna

D. Fase *Construction*

Fase ini juga dikenal sebagai tahap pembangunan, tahap ini melengkapi pembangunan sistem aplikasi yang akan dibangun. Seperti halnya SDLC yang juga membangun aplikasi yang sudah dirancang.

E. Fase *Cutover*

Fase ini menyerupai tugas akhir dalam tahap implementasi SDLC, termasuk konversi data, pengujian.

F. Ukuran kesuksesan website

Untuk mengetahui apakah *website* sudah memenuhi kebutuhan dan berjalan dengan baik, keberhasilan website ini ditentukan dengan ukuran dengan tingkat yang dibuat seperti berikut :

- Dapat mengolah data dari ban-pt.kemdiknas.go.id secara otomatis dengan penerapan *web scrapping*.
- Aplikasi mempunyai data yang valid dan akurasi yang baik dalam memberikan informasi kepada pengguna.
- Dapat menampilkan data pada tabel yang memberikan fungsi *search* memberikan output yang sesuai dengan pengguna.
- Menampilkan rasio

IV. ANALISIS DAN PEMBAHASAN

A. Fase *Requirements Planning*

Dari hasil *requirements Planning* yang dilakukan dalam pembuatan webisite dapat disimpulkan bahwa kebutuhan yang diperlukan dalam pembuatan website adalah sebagai berikut:

- Aplikasi berbasis *web application*, yang bertujuan untuk mempermudah pengguna dalam mengakses.
- Aplikasi dapat menyediakan data informasi yang valid sesuai dengan data-data perguran tinggi dan data sensus penduduk yang dikeluarkan oleh Ban-pt dan data.go.id.
- Pengguna dapat melakukan pencarian perguruan tinggi dengan melalui aplikasi yang akan dibangun.

B. Fase *User Design*

Pada fase *user design* dirancang pada halaman awal atau index dan halaman pencarian atau halaman *Search_Page* sesuai dengan perancangan website sebagai berikut:

- Pada halaman *index* website, pengguna akan diberikan beberapa menu dalam halaman awal..
- Pada halaman *index website*, pengguna akan diberikan suatu fungsi pencarian.
- Pada halaman pencarian pengguna dapat melihat tabel dan fungsi pencarian untuk melakukan pencarian kembali.
- Pada halaman pencarian pengguna melihat pada *row* pertama sebagai informasi jumlah total penduduk, berdasarkan dari hasil katagori yang dipilih.
- Pada tabel *data* yang ditunjukan, pengguna dapat melakukan fungsi sort pada tabel.

C. Fase *Construction*

Proses lengkap dari setiap tahap yang dilewati dalam pengelolaan data, dari data kotor hingga menjadi data bersih yang ditampilkan di *website* adalah sebagai berikut:

1) Pencarian Data

Pencarian data dilakukan dari *website* terkait seperti BAN-PT, Dikti, dan data.go.id untuk diabil datanya. Pada tahap pencarian, data yang didapatkan berbentuk URL (*uniform resource locator*) untuk dimasukan dalam pengambilan data.

2) Pengambilan Data

Pengambilan data dilakukan dengan cara *Web Scrapping* dengan menggunakan Tools *Pentaho Data Integration,* yang akan memalu beberapa tahap yaitu:

- *Text File input* proses memasukan data pada *text file* yang berisikan *source code* dalam bahasa HTML yang di ambil dari http://ban-pt.kemdiknas.go.id/ untuk mengambil data
- *Filter rows* digunakan untuk menyaring data yang akan digunakan untuk penelitian ini
- *Replacement string* untuk menghilangkan data-data yang tidak di perlukan, agar hasil yang diterima menjadi data bersih
- *Row Flattener* digunakan untuk pengurutan data yang di ambil di ambil

3) Data Cleansing

Data dibersihkan menggunakan *Pentaho Data Integration* dengan dependensi pada nomor kopertis sehingga tidak akan keluar dari wilayah kopertisnya. Data yang dupilkat akan dibersihkan dengan nama yang disamakan. Berikutnya mengidentifisikasi nomor kode kota/kabupaten yang ada pada setiap perguruan tinggi agar dapat digabungkan pada data penduduk yang ada sehingga dapat di identifikasi berapa jumlah penduduk yang ada.

4) Penggabungan Data

Penggabungan data dalam *Pentaho Data Integration* ditujukan untuk menggabungkan dua data yang berbeda. Dengan fungsi "*Join row*" pada *Pentaho Data Integration* kedua data dapat menjadi satu dan mempunyai dependensi yang sama.

D. Rasio

Rasio digunakan untuk mengetahui kesenjangan edukasi pada perguruan tinggi dari setiap daerah Indonesia. Penelitian ini menggunakan *Gross Enrolment Ratio* (GER). Untuk menghasilkan data yang tepat digunakan data terakhir dari sensus penduduk dan data

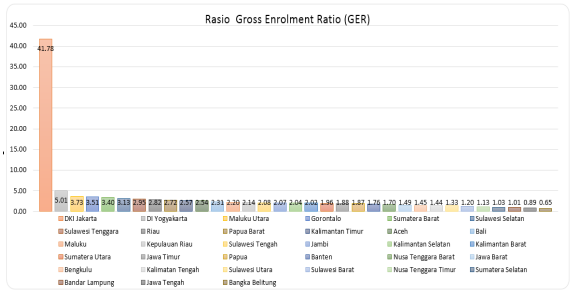
DIKTI untuk digunakan dalam perhitungan rasio.

$$GER_h^t = \frac{E_h^t}{P_{h,a}^t} * 100 \quad (1)$$

dimana

- $GER_h^t GER_h^t$ adalah *Gross Enrolment Ratio* pada level edukasi **h** dalam tahun sekolah **t**
- $E_h^t E_h^t$ adalah Pendaftaran pada level edukasi **h** dalam tahun sekolah **t**
- $P_{h,a}^t P_{h,a}^t$ adalah Populasi dalam grup usia **a** dimana koresponden pada level edukasi **h** dalam tahun sekolah **t**

Gross Enrolment Ratio yang digunakan untuk tingkat umum partisipasi dalam tingkat pendidikan, dan dapat menjadi indicator melengkapi angka partisipasi murni.. Nilai yang mendekati atau melebihi 100% menunjukkan bahwa suatu daerah mampu menampung golongan usia sekolah, tetapi tidak menunjukan proporsi yang sudah terdaftar. Oleh karena itu pencapaian 100% adalah kondisi yang diperlukan tetapi tidak dapat menjadi suatu indikasi yang pasti. Jika sudah mendekati 90% kondisi merupakan indikasi menunjukkan bahwa suatu daerah mampu menampung golongan usia mahasiswa.



Gambar 5. Hasil perhitungan rasio GER per provinsi di Indonesia

E. Fase *Cutover*

Website akan diuji setelah dibangun untuk mengukur apakah *website* yang dibuat dapat memberikan informasi yang sesuai dengan ekspetasi penulis dan memastikan website bekerja dengan baik. Pengujian yang akan dilakukan adalah keakurasian data dan validitas data yang diberikan oleh website.

Pengukuran validitas dan akurasi pada *website,* dengan cara membandingkan dengan data yang dapatkan pada ban-pt.kemdiknas.go.id dan data.go.id.

Tabel 1. Sampling pengujian validasi data perguruan tinggi

perguruan_tinggi	database			website		
	Program studi	Daluwarsa	status	Program studi	Daluwarsa	status
Institut Teknologi Bandung, Bandung	Aeronautika dan Astronautika	27-08-2015	masih berlaku	Aeronautika dan Astronautika	27-08-2015	masih berlaku
Universitas Kristen Satya Wacana, Salatiga	Ilmu Komunikasi	18-10-2017	masih berlaku	Ilmu Komunikasi	18-10-2017	masih berlaku
Universitas Indonesia, Jakarta	Administrasi Asuransi dan Aktuaria	9/6/2016	masih berlaku	Administrasi Asuransi dan Aktuaria	9/6/2016	masih berlaku
Sekolah Tinggi Ilmu Administrasi dan Sekretari Bunda Hati Kudus	Administrasi Bisnis	13-08-2014	Kadaluarsa	Administrasi Bisnis	13-08-2014	Kadaluarsa
Sekolah Tinggi Ilmu Administrasi Amal Ilmiah Wamena	Administrasi Negara	2/4/2014	Kadaluarsa	Administrasi Negara	2/4/2014	Kadaluarsa

Pengujian juga mencakup validitas pada jumlah populasi pada provinsi dan daerah dengan cara yang sama sama. Pengujian ini bertujuan untuk mengetahui apakah sistem dalam penjumlah data telah berjalan dengan baik

Tabel 2. Sampling pengujian validasi data penduduk

Wilayah	database				website			
	populasi	umur 18	umur 19	umur 20-24	populasi	umur 18	umur 19	umur 20-24
ACEH	610965	85225	87344	438396	610965	85225	87344	438396
kab.Sumedang	122988	19295	20122	83571	122988	19295	20122	83571
kota semarang	333741	47530	52477	233743	333741	47530	52477	233743
jakarta	1368122	173765	183958	1010399	1368122	173765	183958	1010399
sulawesi tenggara	278309	41590	39091	197628	278309	41590	39091	197628

Dengan melakukan data sampling dapat membuktikan bahwa aplikasi dapat menghasilkan informasi perguruan tinggi dengan valid. Kedua pengujian yang telah dilakukan pada kedua pengujian diatas dapat memenuhi ekspetasi. Tingkat validitas data dan akurasi telah memenuhi kesuksesan *website*.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

- *Website* ini dibangun untuk mengetahui gambaran kesenjangan edukasi pada tingkat perguruan tinggi setiap daerah yang ada di Indonesia. Sistem pengumpulan data pada *website* dibangun dengan cara *web crawling, web scrapping,* dan *data cleansing* untuk

mendapatkanan data-data yang valid

- Pada perhitungan *Gross Enrolment Ratio* (GER) yang didapatkan dari penelitian ini, Indonesia mendapatkan hasil 2,03%, dan untuk perhitungan rasio per daerah, DKI Jakarta mendapatkan rasio tertinggi senilai 41,78%. Dari hasil rasio yang didapatkan bisa ditarik kesimpulan bahwa Indonesia belum dapat menampung golongan usia mahasiswa dari standar yang diberikan oleh UNESCO.
- *Website* yang dibangun telah berhasil menunjukan rasio, dapat mengolah data, dapat menampilkan data tabel pencarian, dan dapat memberikan data dan informasi yang valid serta akurasi yang baik.

B. Saran

- Agar mendapatkan data yang lebih aktul dan terbaru dari UNESCO.
- Agar dapat memberikan gambaran secara keseluruan tentang pendidikan di Indonesia, *website* dapat dikembangakan dengan cara menambahkan data akreditasi sekolah, perwilayah.
- Terkait dengan pengembangan *website* ini, agar dapat memberikan tambilan hasil *search* lebih menarik dan mudah dimengerit pengguna

Ucapan Terima Kasih

Penulis mengucapkan terimakasih kepada Bapak Wira Munggana, S. Si., M.Sc. dan Ir. Raymond Sunardi Oetama, M.C.I.S yang memberikan dukungan, saran dan inspirasi dalam penelitian ini.

DAFTAR PUSTAKA

[1] Cambridge, "Viral," Cambridge Dictionaries, [Online]. Available: http://dictionary.cambridge.org/dictionary/british/viral. [Accessed 2015 Maret 3].
[2] Franklin & Marshall College, "Data Collection," Franklin & Marshall College, [Online]. Available: http://www.fandm.edu/opinionresearch/data-collection. [Accessed 20 Maret 2015].
[3] O. Christopher and N. Marc, Web Crawling, California: standford, 2010.
[4] OLAP, "Business Intelligence," PARIS Technologies, [Online]. Available: http://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/. [Accessed 3 Maret 2015].
[5] T. Matthew, php|architect's Guide to Web Scrapping with PHP, Alexandria: musketeers.me, LLC, 2010.
[6] UNESCO, "UNESCO Institute for Statistics," [Online]. Available: http://www.uis.unesco.org/Library/Documents/eiguide09-en.pdf. [Accessed 13 Mei 2015].
[7] University of Wisconsin Eau Claire, "Data Collection Methods," University of Wisconsin Eau Claire, [Online]. Available: http://people.uwec.edu/piercech/researchmethods/data%20collection%20methods/data%20collection%20methods.htm. [Accessed 5 Maret 2010].