Web Scraping Dengan *HTML DOM* Untuk Pengumpulan Data NPSN Dari *Website* Data Referensi Kemendikbudristek (Studi Kasus Pada Situs PMB STT POMOSDA Nganjuk)

Ibnu Sodik[1)], Anang Efendi[2)], Gendut Sadar Laswijiyanto[3)], Muhammad Jamal[4]

[1,2,3)]Program Studi Teknik Informatika, STT Pomosda Nganjuk
e-mail: [1]ibnusodik049@gmail.com, [2]afendystt@gmail.com, [3]laswijiyanto07@gmail.com, [4]jamal@st-pomosda.ac.id

Abstract

As a higher education institution, it is our obligation to provide a registration system to make it easier for users to register online. In efforts to develop a online new student registration system, a reference source is needed as a reference to make it easier for users to fill out the registration form. Web scraping is a technique or process used to retrieve information or data from a website automatically and use it for certain purposes. In this context, web scraping is used to collect NPSN (National School Principal Number) data from the Kemendikbudristek (Ministry of Education and Culture, Research and Technology) website. To solve this problem, researchers used web scraping to collect npsn data combined with HTML DOM which was applied to the registration form for the New Student Registration system at Pondok Modern Sumber At-Taqwa Technology College. The method used is the Rapid Application Development (RAD) method which focuses on a fast system development process. This research succeeded in developing a new student registration system, used scraped data to be displayed on the registration form using HTML DOM, and succeeded in collecting all open data for the secondary education category with a total of 41,545 data on the Ministry of Education and Culture's reference data site and succeeded in saving the scraped data in the form databases.

Keywords: Web Scraping, HTML DOM, NPSN

Abstrak

Sebagai salah satu lembaga pendidikan tinggi sudah menjadi kewajiban menyediakan sebuah sistem pendaftaran untuk memudahkan pengguna saat akan melakukan pendaftaran secara *online*. Dalam upaya pengembangan sistem pendaftaran mahasiswa baru secara *online* dibutuhkan sumber referensi sebagai rujukan agar memudahkan pengguna saat melakukan pengisian *form* pendaftaran. *Web scraping* adalah teknik atau proses yang digunakan untuk keperluan pengambilan informasi atau data dari sebuah *website* secara otomatis dan digunakan untuk keperluan tertentu. Dalam konteks ini, *web scraping* digunakan untuk mengumpulkan data NPSN (Nomor Pokok Sekolah Nasional) dari *website* Kemendikbudristek (Kementerian Pendidikan dan Kebudayaan, Riset, dan Teknologi). Untuk menyelesaikan masalah tersebut, peneliti menggunakan *web scraping* untuk mengumpulkan data npsn digabungkan dengan *HTML DOM* yang diterapkan pada *form* pendaftaran sistem Pendaftaran Mahasiswa Baru Sekolah Tinggi Teknologi Pondok Modern Sumber Daya At-Taqwa. Metode yang digunakan adalah metode *Rapid Application Development (RAD)* yang berfokus kepada proses pengembangan sistem yang cepat. Penelitian ini berhasil mengembangkan sistem pendaftaran mahasiswa baru, menggunakan data hasil *scraping* untuk ditampilkan pada *form* pendaftaran menggunakan *HTML DOM*, berhasil mengumpulkan semua data npsn kategori pendidikan menengah dengan jumlah total 41545 data yang ada pada situs data referensi kemendikbudristek dan berhasil menyimpan data hasil scraping kedalam bentuk *database*.

Kata Kunci : *Web Scraping, HTML DOM, NPSN*

Pendahuluan

Data Pokok Pendidikan, yang selanjutnya disingkat Dapodik adalah salah satu sistem pendataan yang dikelola oleh Kementrian Pendidikan dan Kebudayaan yang memuat data satuan pendidikan, peserta didik, pendidik dan tenaga kependidikan, dan substansi pendidikan yang datanya bersumber dari satuan pendidikan yang terus menerus diperbarui secara online (Jasuma dkk., 2019). Sistem Dapodik dirancang menggunakan basis *open source* dengan menerapkan sistem *database* terpusat dan aplikasi berbasis web. Dengan sistem tersebut maka pengelolaan riwayat data sekolah, siswa, guru/karyawan lebih mudah terintegrasikan dan disimpan secara terpusat sehingga dapat diakses dengan mudah oleh masyarakat melalui internet (Tueno dkk., 2020).

Web Scraping adalah teknik untuk mendapatkan informasi dari *website* secara otomatis tanpa harus menyalinnya secara manual (Deviacita dkk., 2019). *Web scraping* bekerja dengan melakukan proses ekstrasi data dengan mempelajari kode dari sebuah *website* yang hendak diambil data informasinya, data yang ingin ditarik biasanya berbentuk teks bertipe HTML atau XHTML (Fauzia Putri dkk., 2021). Hasil *web scraping* pada *website* data referensi kemendikbudristek akan digunakan untuk membuat layanan data NPSN berupa file *database* dan menguji *database* yang telah dibuat untuk mengetahui apakah *database* tersebut cukup baik untuk digunakan. Pada penelitian yang dilakukan oleh Firdian dkk. (2022) metode *HTML DOM* digunakan dalam penelitian ini karena bedasarkan penelitian sebelumnya metode tersebut memiliki rata-rata waktu *scraping* lebih cepat dan menggunakan internet lebih sdikit dibandingkan dengan dua metode lainnya yaitu *Reguler Expression* dan *Xpath*.

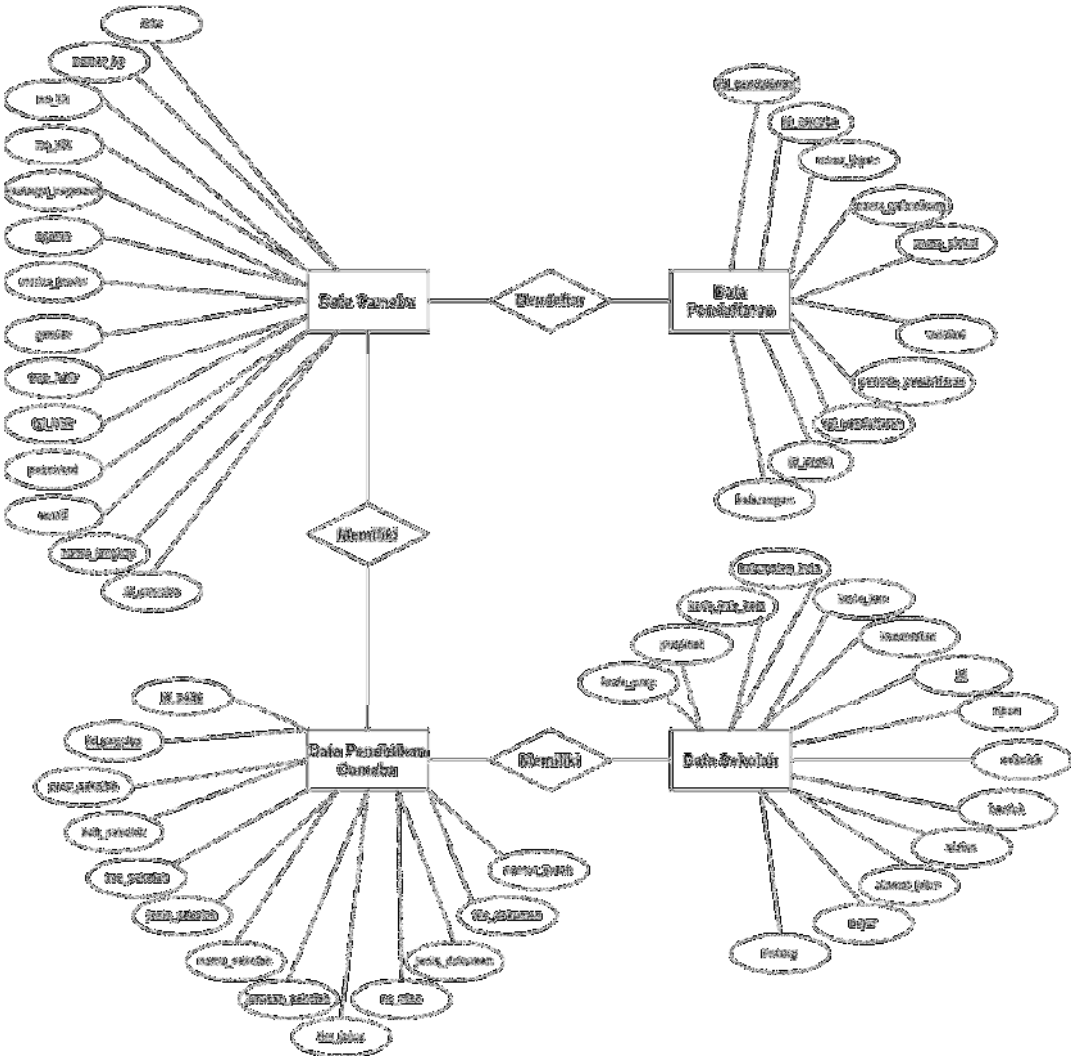
Berdasarkan latar belakang penelitian ini maka rumusan masalah yang dapat diambil adalah a) Bagaimana cara mengumpulkan data NPSN dari *website* kemendikbudristek menggunakan teknik *web scraping*. b) Bagaimana penerapan *HTML DOM* pada *web scraping* untuk pengumpulan data NPSN dari *website* kemendikbudristek. c) Apakah hasil *scraping* data NPSN dari *website* kemendikbudristek dapat digunakan kembali. Adapun tujuan penelitian ini adalah a). Mengumpulkan data NPSN dari website referensi kemendikbudristek menggunakan teknik web scraping. b) Menerapkan HTML DOM pada web scraping untuk pengumpulan data NPSN dari website kemendikbudristek. c) Menggunakan kembali data NPSN hasil scraping dari website kemendikbudristek.

Analisa dan Perancangan Sistem

Analisis kebutuhan sistem memuat spesifikasi yang rinci tentang hal yang akan dilakukan sistem ketika diimplementasikan sesuai dengan kebutuhan pengguna sistem. Pada tahap ini dilakukan identifikasi terhadap situs *website* data referensi kemendikbudristek sebagai pusat data npsn seluruh wilayah Indonesia. Data npsn tersebut akan dikumpulkan menjadi sebuah *database* yang akan diterapkan hasilnya pada sistem PMB STT POMOSDA.

Data npsn tersebut akan dikumpulkan menggunakan teknik *web scraping* dan akan ditampikan hasilnya kedalam *form* pendaftaran bagian asal sekolah menggunakan *HTML DOM,* sehingga pengguna atau pendaftar calon mahasiswa baru dapat lebih mudah dalam mengisi *form* pendaftaran

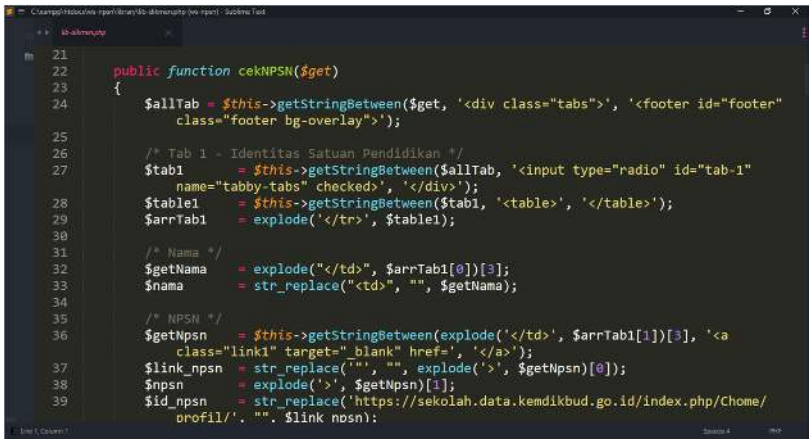
ERD atau *Entity Relationship Diagram* adalah suatu gambaran struktural yang digunakan dalam perancangan sebuah *database*. Diagram *ER* digunakan untuk menggambarkan data yang akan disimpan dalam suatu sistem serta batas-batasnya (Larassati dkk., 2019).



Gambar 3. 1 *ERD* Sistem PMB

Hasil Dan Pembahasan

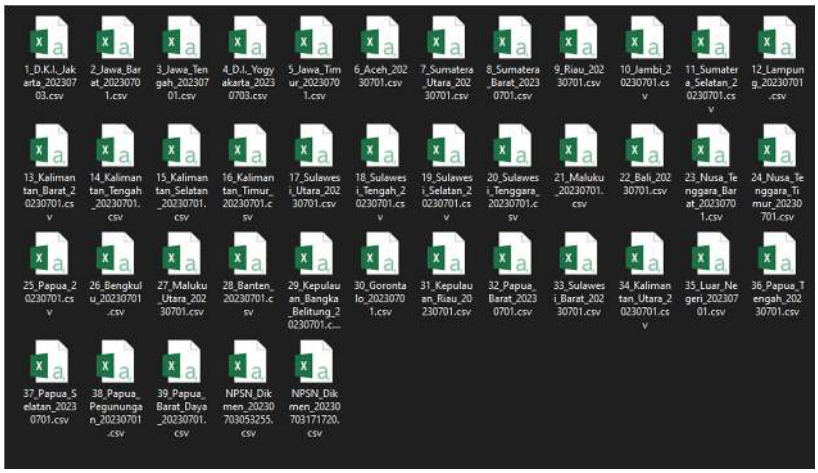
Implementasi *web scraping* dilakukan dengan tahapan identifikasi kerangka situs *web*, membuat *script web scraping*, *scraping* data, hasil *scraping,* membuat file *csv* dan input *database***.** Data yang akan diekstrak terdiri dari data provinsi, data kabupaten, data kecamatan, data kelurahan, data NPSN dan detail NPSN. Masing-masing data tersebut memiliki keterkaitan anatara satu sama lain, oleh karena itu perlu dilakukan identifikasi *website* terlebih dahulu pada situs kemendikbud. Menurut Firdian dkk. (2022) tahapan ini dilakukan sebagai acuan dalam membuat *script* untuk *scraping*.



```
21
22      public function cekNPSN($get)
23      {
24          $allTab = $this->getStringBetween($get, '<div class="tabs">', '<footer id="footer"
                class="footer bg-overlay">');
25
26          /* Tab 1 - Identitas Satuan Pendidikan */
27          $tab1       = $this->getStringBetween($allTab, '<input type="radio" id="tab-1"
                name="tabby-tabs" checked>', '</div>');
28          $table1     = $this->getStringBetween($tab1, '<table>', '</table>');
29          $arrTab1    = explode('</tr>', $table1);
30
31          /* Nama */
32          $getNama    = explode("</td>", $arrTab1[0])[3];
33          $nama       = str_replace("<td>", "", $getNama);
34
35          /* NPSN */
36          $getNpsn    = $this->getStringBetween(explode('</td>', $arrTab1[1])[3], '<a
                class="link1" target="_blank" href=', '</a>');
37          $link_npsn  = str_replace('"', "", explode('>', $getNpsn)[0]);
38          $npsn       = explode('>', $getNpsn)[1];
39          $id_npsn    = str_replace('https://sekolah.data.kemdikbud.go.id/index.php/Chome/
                profil/', "", $link_npsn);
```

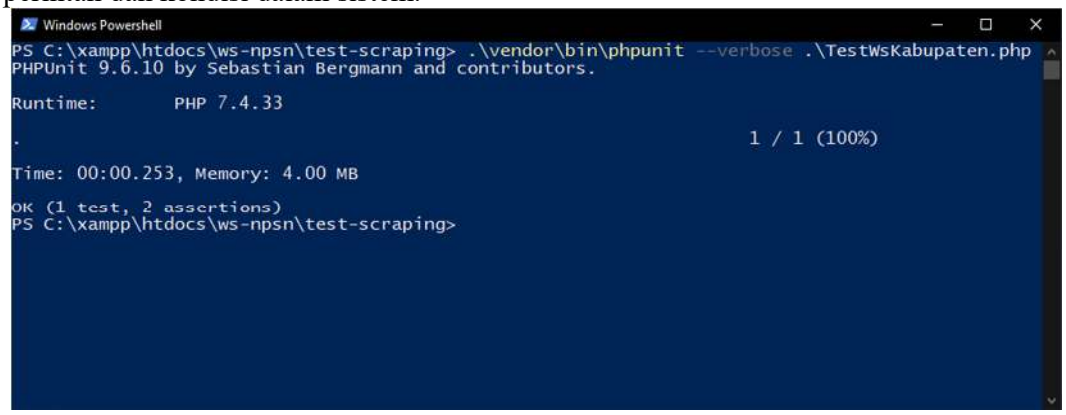
Gambar 4. 1 *Script Web Scraping*

Gambar 4. 2 Hasil Scraping

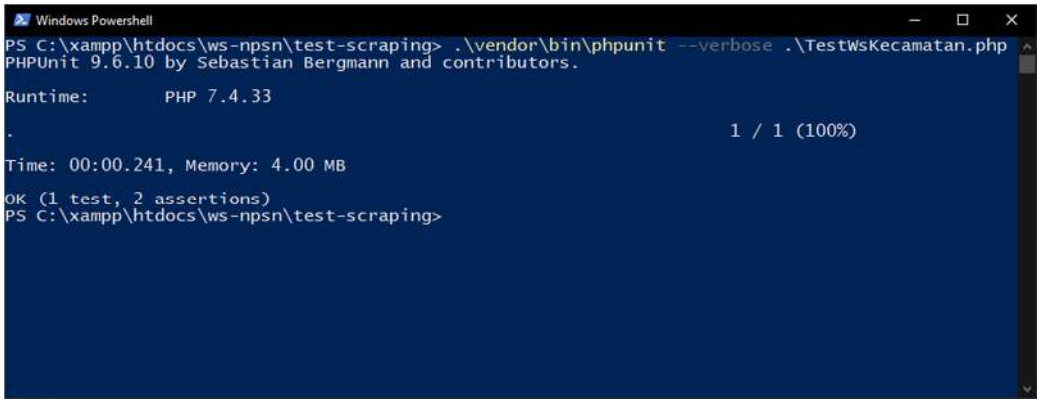


Gambar 4. 3 Data *Scraping* Dalam Bentuk File *Csv*

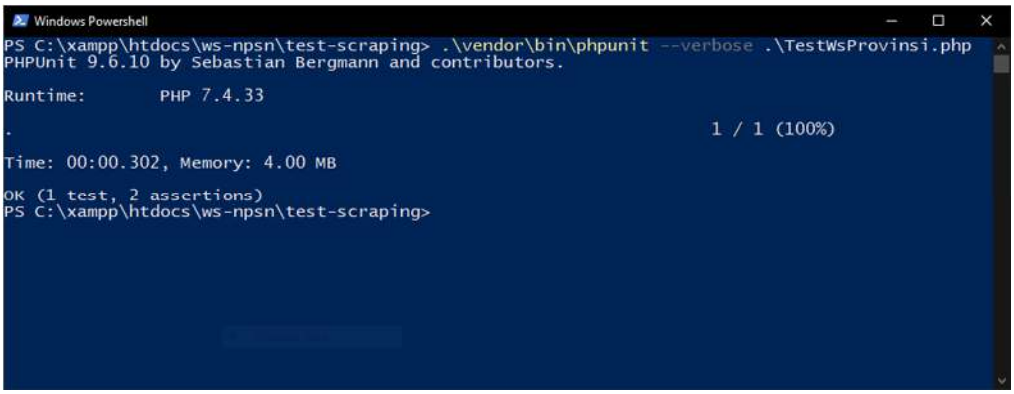
Pengujian *white box testing* memiliki tujuan untuk memverifikasi eksekusi semua perintah dan kondisi dalam sistem.



Gambar 4. 4 *Scraping* data NPSN berdasarkan Provinsi



Gambar 4. 5 *Scraping* data NPSN berdasarkan Kabupaten

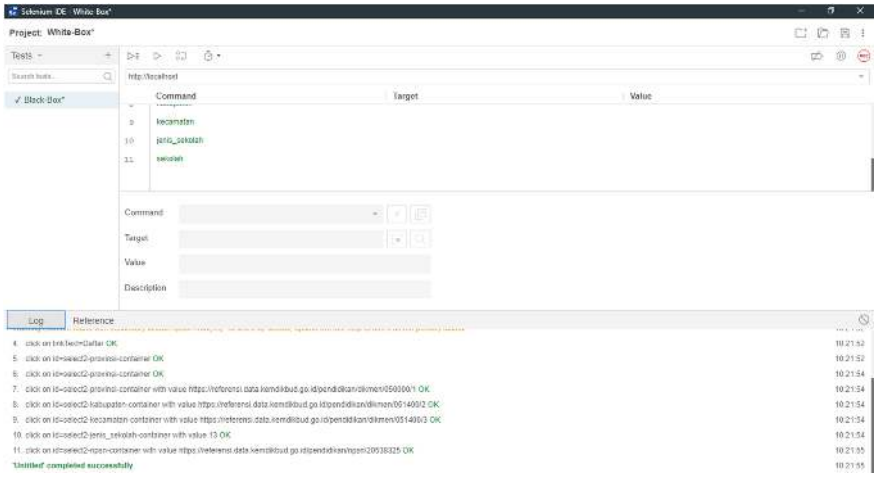


Gambar 4. 6 Scraping data NPSN berdasarkan Kecamatan

Pengujian terhadap fitur-fitur dalam program Web Scraping Dengan HTML DOM Untuk Pengumpulan Data NPSN Dari Website Kemendikbudristek menggunakan metode black box

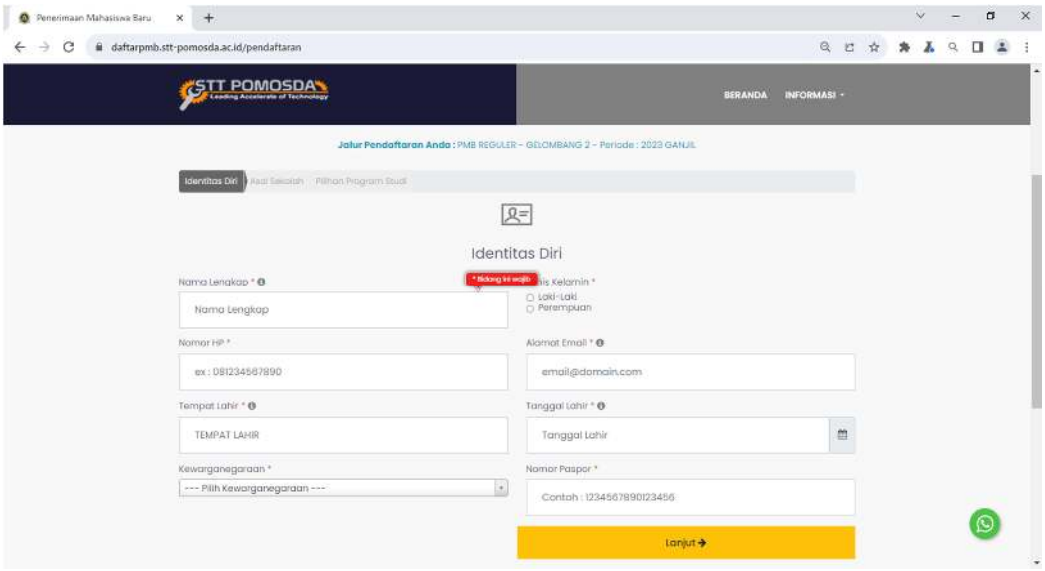
testing, proses ini dilakukan untuk memeriksa apakah program yang dibangun berjalan dengan benar sesuai dengan yang diharapkan atau tidak (Deviacita dkk., 2019).

Data untuk pengujian menggunakan metode blackbox testing ini dipilih berdasarkan keluaran program yang diharapkan tanpa memperhatikan detail internal dari program. Hasil pengujian dapat dilihat pada gambar 4.13.

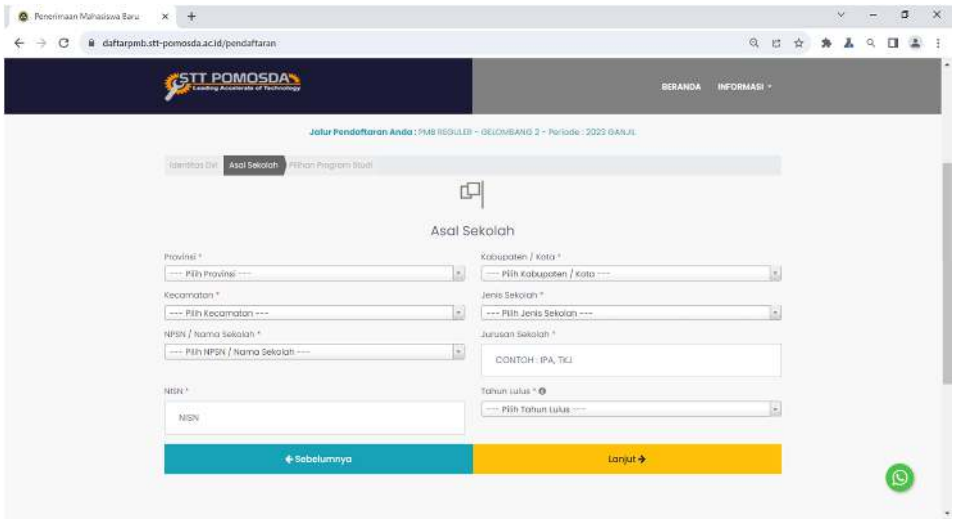


Gambar 4. 7 Hasil Pengujian Sistem

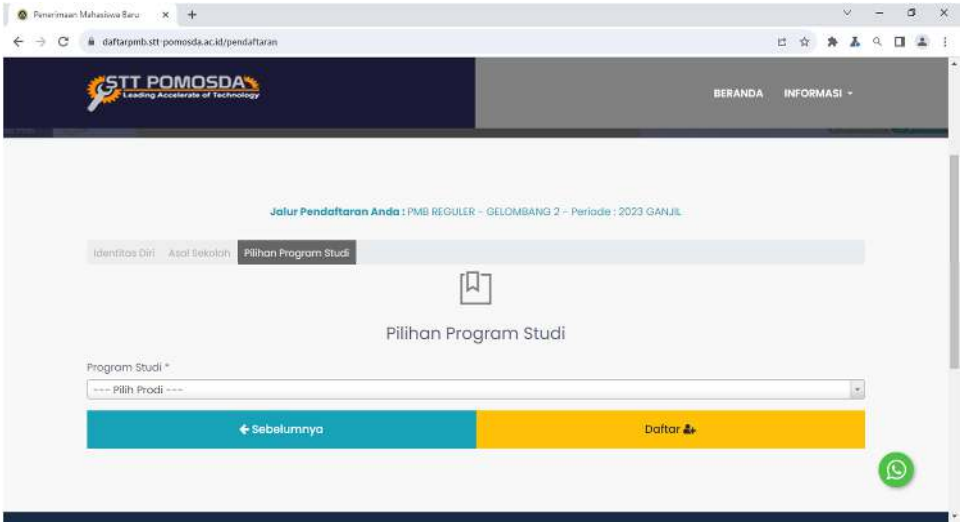
Pengujian pertama menggunakan Chromium, menampilkan form pendaftaran untuk isian data identitas diri, asal sekolah dan pilihan prodi. Pada form bagian asal sekolah berhasil menampilkan data hasil scraping data npsn. Hasil pengujian dapat dilihat pada gambar 4.10, 4.11 dan 4.12.



Gambar 4. 8 Pengujian Form Identitas Diri Menggunakan Chromium



Gambar 4. 9 Pengujian Form Asal Sekolah Menggunakan *Chroium*



Gambar 4. 10 Pengujian Form Pilihan Prodi Menggunakan *Chromium*

Penutup

Dari hasil penelitian yang dilakukan dapat disimpulkan sebagai berikut.

1. Telah dikembangkan sistem PMB STT POMOSDA dengan tampilan baru, *HTML DOM* berhasil diterapkan untuk menampilkan data npsn pada form pendaftaran bagian asal sekolah untuk mempermudah pengguna saat melakukan pengisian data pendaftaran.
2. *Web scraping* berhasil mengumpulkan data npsn kategori pendidikan menengah sebanyak 41545 dari 41557 data npsn yang ada pada *website* data referensi kemendikbudristek, menyimpan data hasil scraping dalam bentuk file *csv*, dan berhasil mengimpor data hasil scraping kedalam *database*.
3. Terdapat data npsn yang tidak dapat diambil oleh *web scraping* yaitu data peta karena terjadi permasalahan pada simbol petik satu (‘) dan petik dua (“) yang mempengaruhi proses impor data ke *database*. Akan tetapi masalah tersebut dapat diatasi karena *web scraping* dapat mengambil data bujur dan lintang dari detail npsn sehingga data tersebut dapat menggantikan data peta..

Saran

Berdasarkan perancangan dan hasil implementasi sistem yang dilakukan, maka beberapa saran yang perlu diperhatikan dalam pengembangan sistem ini adalah Mengembangkan sistem PMB STT POMOSDA khususnya pada form pendaftaran dengan hanya memasukkan NIK maka form identitas diri terisi otomatis.

Daftar Pustaka

Deviacita, D., #1, A., Sasty, H., #2, P., Muhardi, H., Profesor, J., Nawawi, D. H. H., Laut, B., Tenggara, P., Pontianak, K., & Barat, K. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal SIstem dan Teknologi Informasi*, *7*(4).

Fauzia Putri, A., Manik, G., Nabila, F., & Chamidah, N. (2021). *Implementasi Scraping Google Scholar Menggunakan HTML DOM Untuk Pengumpulan Data Artikel Dosen UPN Veteran Jakarta Berbasis Web*.

Firdian, M. I., Darwiyanto, E., & Adrian, M. (2022). *WEB SCRAPING DENGAN METODE HTML DOM UNTUK SITUS WEB PEMBUATAN API BERITA , Maulana Irfan Firdian1) Eko Darwiyanto2) , dan Monterico Adrian3)*.

Jasuma, A., Wijayanti, R., Febriani, S., Putro, S., Wisnu, S., Yulia, E., & Yudano, A. (2019). Analisis Data Dapodik Pada SMA ABC di Yogyakarta Sebagai Bagian Evaluasi Sekolah. *Julyxxxx*, *x, No.x*, 1–5.

Larassati, M., Latukolan, A., Arwan, A., & Ananta, M. T. (2019). Pengembangan Sistem Pemetaan Otomatis Entity Relationship Diagram Ke Dalam Database. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(4), 4058–4065. http://j-ptiik.ub.ac.id

Tueno, N. S., Tinggi, S., Bina, I. A., & Gorontalo, T. (2020). FAKTOR-FAKTOR PENGHAMBAT PELAKSANAAN SISTEM APLIKASI DAPODIK DALAM PEMBAYARAN TUNJANGAN PROFESI GURU DI SMP NEGERI 2 KWANDANG. *UBLIK: Jurnal Manajemen Sumber Daya Manusia, Adminsitrasi dan Pelayanan Publik*, *7*(1).