Jurnal Siliwangi Vol.7. No.2, 2021 **P-ISSN 2477-3891**
Seri Sains dan Teknologi **E-ISSN 2615-4765**

WEB SCRAPING SITUS E-COMMERCE MENGGUNAKAN TEKNIK PARSING DOM

Farhan Djiwadikusumah[1)], Genta Hayindra Irawan[2)], Rifqy Haekal al-Fadilah[3)]

e-mail: 177006027@student.unsil.ac.id[1], 177006028@student.unsil.ac.id[2], 177006076@student,unsil.ac.id[3]
Program Studi Informatika Universitas Siliwangi
Jl. Siliwangi No.24, Kahuripan, Kec. Tawang, Tasikmalaya, Jawa Barat 46115

Abstrak

Kegiatan jual-beli eletronik atau *e-commerce* melibatkan transfer dana dan transfer data secara eletronik, sistem manajemen inventori dan sistem pengumpulan data otomatis dalam melakukan pemasaran, penjualan, dan pembelian barang dan jasa. Terdapat banyak perusahaan yang bergerak dalam bidang *e-commerce* diantaranya: Tokopedia, Shopee, JD.in, elevenia, blibli, dan Bukalapak. Setiap perusahaan ini menyediakan situs "pasar online" yang dapat mencari, memilih, dan membeli produk yang diinginkan pembeli. Demi menentukan keputusan terbaik dalam pembelanjaan, perlu dilakukan pencarian di beberapa situs jual beli untuk melihat kualitas toko dan pelayanannya, dan tingkat kepopuleran barang. Guna mempermudah dan mempercepat pembelian, dapat ditangani dengan menggunakan situs yang dapat menampilkan lebih dari satu situs toko e-commerce agar user bisa dengan mudah membandingkan antar situs. Salah satu caranya menggunakan teknik web scraping. Dengan web scraping dapat dilakukan pengambilan data dengan cepat, akurat, dan otomatis dari website target. Tujuan dari penelitian ini, menerapkan metode parsing DOM pada web scraping situs e-commerce. Berdasarkan hasil pengujian, sistem mampu memberikan hasil sesuai harapan awal dari tiga situs e-commerce.

Kata Kunci : Web Scraping, E-Commerce, Parsing DOM

Abstract

Electronic buying and selling activities or e-commerce involve electronic transfers of funds and data, inventory management systems and automated data collection systems in marketing, selling, and purchasing goods and services. There are many companies engaged in e-commerce including: Tokopedia, Shopee, JD.in, elevenia, blibli, and Bukalapak. Each of these companies provides "online marketplace" sites where you can search, select, and buy the products that shoppers want. In order to determine the best decision in shopping, it is necessary to do a search on several buying and selling sites to see the quality of the store and its services, and the level of popularity of the goods. In order to simplify and speed up purchases, it can be handled by using sites that can display more than one e-commerce store site so that users can easily compare between sites. One way is to use web scraping techniques. With web scraping, data can be retrieved quickly, accurately, and automatically from the target website. The purpose of this study is to apply the DOM parsing method on web scraping e-commerce sites. Based on the test results, the system is able to provide results according to the initial expectations of the three e-commerce sites.

Keywords*: Web Scraping, E-Commerce, Parsing DOM*

I. PENDAHULUAN

Teknologi internet berkembang dengan pesat, mudah diakses dan digunakan oleh masyarakat. Pengguna internet dengan mudah mendapat informasi untuk keperluan aktivitas sehari-hari, mencari informasi berbagai barang atau jasa yang dibutuhkan. Bagi para penggiat bisnis hal, ini merupakan peluang besar untuk memperluas jangkauan bisnis, sehingga bermunculan banyak toko *online*. Toko *online* dapat melakukan penjualan secara independen, melalui perantara forum seperti forum jual beli kaskus. Selain dari itu juga dapat melakukan transaksi penjulan melalui perusahaan digital seperti Tokopedia, Blanja, Elevenia, dan lainnya. Agar mempermudah pengguna untuk mencari produk yang diinginkan, dengan hasil penjualan terbaik, serta perbandingan harga produk antar situs jual beli komersial, maka diusulkan suatu media yang dapat mengumpulkan berbagai informasi yang terdapat pada toko komersial online yang akan menampilkan informasi penting berupa barangnya, harga, nama produk, dan asal situsnya dengan cara *web scrapping*. *Web scrapping* merupakan suatu proses pembacaan, pengambilan, pengumpulan, atau *crawling* data atau dokumen dari website yang diinginkan, dokumen biasanya berupa. Biasanya dilakukan untuk membantu proses analisa data, menganalisa data kompetitor, melakukan *brand*

monitoring, maupun keperluan SEO karena data yang diambil selalu terbaharukan sesuai perubahan pada paket data yang disediakan oleh situs tersebut. Terdapat beberapa metode dalam pengambilan data (*scraping*) dari situs ini antara lain *copu/paste* dimana dilakukan salin dan tempel manual data yang diinginkan, biasanya dipakai untuk blogger kecil-kecilan, ada juga pengambilan secara otomatis diantaranya : *HTML Parsing, DOM Parsing, Vertical Aggregation, Xpath, Google Sheets, Text Pattern Matching,* dan sebagainya [1,7,10,12, 13,14].

Dalam penelitian ini, *web scraping* dengan metode *DOM parsing* akan diimplementasikan untuk mendapatkan data dari 3 situs toko komersial Elevenia.com, fjb.kaskus.com, dan blanja.com sekaligus untuk dibandingkan harga dan jenis produk dari ketiga situs web komersial ini.

II. TINJAUAN PUSTAKA

A. *Web Scrapping*

Web scrapping adalah teknik untuk mendapatkan informasi dari situs tertentu secara manual yang dilakukan dengan cara menyalin informasi secara manual maupun otomatis. *Web scraping* berfokus pada mendapatkan data yang dilakukan dengan cara pengambilan dan ekstraksi. Manfaatnya agar informasi yang telah disalin dapat disaring sehingga mempermudah melakukan pencarian suatu data dengan ukuran yang bervariasi[3]. Tahapan yang akan dilakukan yaitu :

1. Permintaan url-url yang akan dijadikan target pengambilan data.
2. Menunggu permintaan yang akan diproses oleh server target.
3. Hasil dari *request* yang kirim oleh server (yang diperlukan dalam *scraping* ini dalam bentuk HTML 5 (yang mengandung XHTML5.1)) akan diambil inforasinya untuk ditiru pada aplikasi yang akan dibuat[4,5].
4. Dilakukan Ekstraksi data untuk mengoptimalkan pengambilan data (teks dalam paket HTML) dan penentuan output pada aplikasi[7,12].

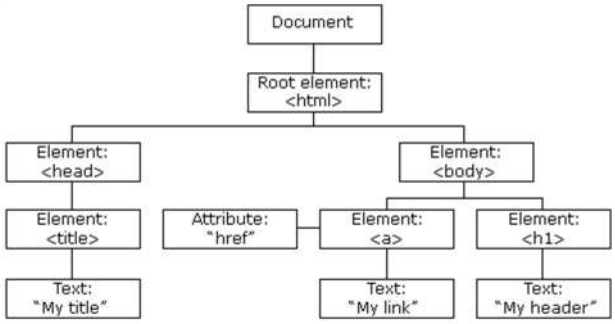
B. *HTML*

HTML merupakan bahasa *markup* (Markah) yang digunakan sebagai dokumen untuk membuat atau mengkonstruksi sebuah halaman web. Terdiri dari elemen-elemen yang mendeskripsikan konten halaman web seperti heading, paragraf, dan list. Dokumen ini yang bertugas dalam menampilkan informasi di sebuah *website*, yang informasi tersebut disimpan dalam format ASCII normal sehingga tertampilkan halaman web dengan perintah-perintah HTML, sebelumnya dimulai dari bahasa SGML (*Standard Generalized Markup Language*)[4,5,9]. Versi HTML terbaru sekarang adalah HTML5. Pada HTML5 terdapat perubahan dari versi sebelumnya dan secara garis besar berusaha untuk merefleksikan kebutuhan dari *website* saat ini dan masa yang akan datang[3, 9].

C. *HTML DOM*

DOM yang merupakan singkatan dari *Document Object Model* merupakan model objek untuk html yang mendefinisikan elemen HTML sebagai objek, properti untuk semua elemen HTML, method untuk semua elemen HTML, dan *event* untuk semua elemen HTML. Dalam JavaScript, HTML DOM merupakan sebuah API (*Programming Interface*), dimana JavaScript dapat menambah, mengubah, maupun menghapus elemen HTML, atribut-atribut HTML, *Style* CSS, dan juga *event* HTML[3, 6].

Disaat suatu laman situs telah dimuat, browser membuat DOM dari laman tersebut. Model HTML DOM pada umumnya disusun sebagai pohon objek *(Tree of Objects*) seperti yang diilustrasikan pada gambar 1.



Gambar 1 HTML DOM Tree of Object

D. *Unified Modeling Language (UML)*

Unified Modeling Language (UML) adalah bahasa spesifikasi standar untuk memvisualisasikan suatu *requirement*, membuat desain, dan analisis di keperluan umum maupun pengembangan (*developmental*) arsitekrur rekayasa perangkat lunak (*software engineering*) dalam pemrograman berorientasi objek[2, 11].

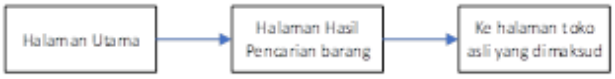
1. *Use Case Diagram*
   Use case diagram menjelaskan tentang interaksi-interaksi antara satu atau lebih aktor dengan sistem informasi yang akan dibuat [8].
2. *Class Diagram*
   Menggambarkan struktur sistem dari segi pendefinisian kelas-kelas yang akan dibuat untuk membuat sistem. Dalam kelas terdapat atribut dan metode/operasi. Atribut disini yaitu variabel-variabel yang dimiliki oleh suatu kelas. Metode/operasi merupakan fungsi-fungsi yang dimiliki oleh suatu kelas.[8,11]

3. *Sequence Diagram*
   Menggambarkan kolaborasi sekumpulan objek yang berinteraksi beserta urutan interaksinya, kegunaannya untuk menunjukkan rangkaian pesan yang dikirim beserta interaksi antar objek [2, 8, 11].
4. *Activity Diagram*
   Diagram aktivitas menggambarkan rangkaian aliar kerja (*workflow*) atau aktifitas sebuah sistem[8, 11].

III. PERANCANGAN APLIKASI

A. Arsitektur Sistem

Desain arsitektur secara garis besar diilustrasikan pada gambar 2.



Gambar 2. Desain Arsitektur

1. Halaman Utama

   Halaman utama ini menampilkan bar dan tombol untuk pencarian produk dengan memanfaatkan fungsi pencarian dari ketiga situs komersial online yang telah dipilih.

2. Halaman Pencarian

   Halaman pencarian akan menampilkan list hasil pencarian barang berdasarkan kata kunci yang telah dimasukkan ke bar pencarian tadi.

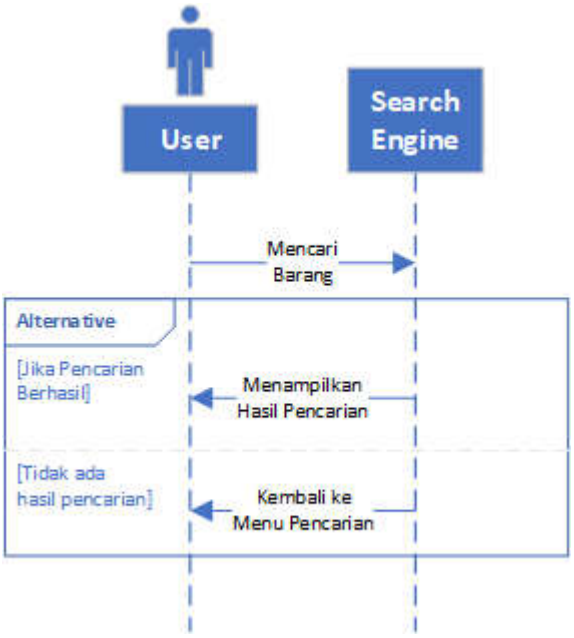
3. *Hyperlink* toko

   Di tiap tiap *item* hasil pencarian terdapat tombol yang mengacu pengguna ke situs aslinya.

B. *Use Case Diagram*

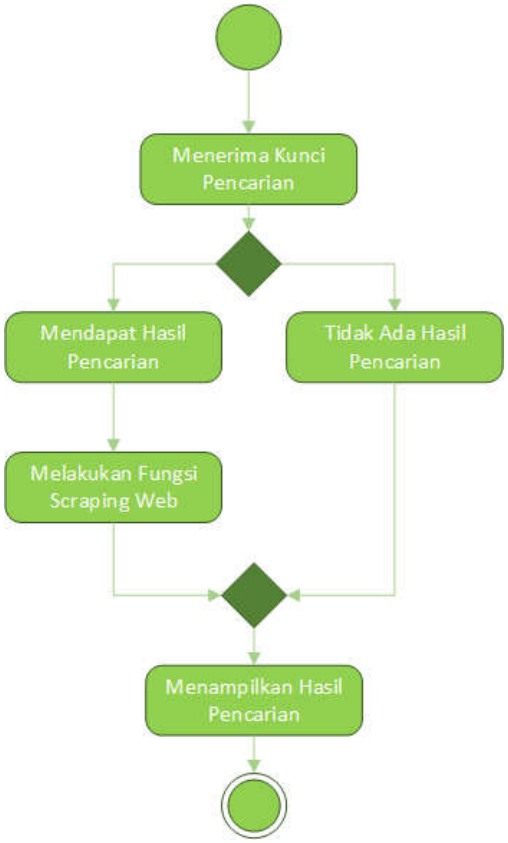
Gambar 3. Use Case Diagram

C. *Sequence Diagram*



Gambar 4. Sequence Diagram

D. *Activity Diagram*

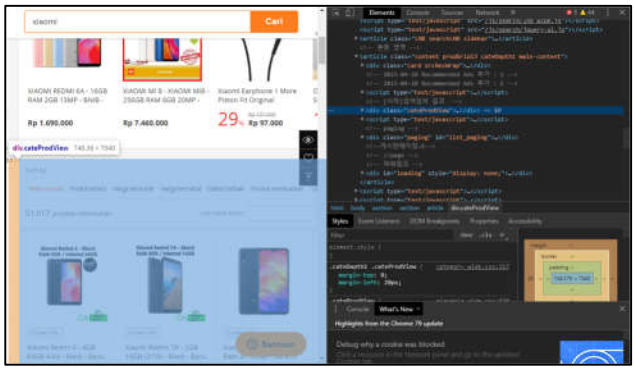


Gambar 5. Activity Diagram

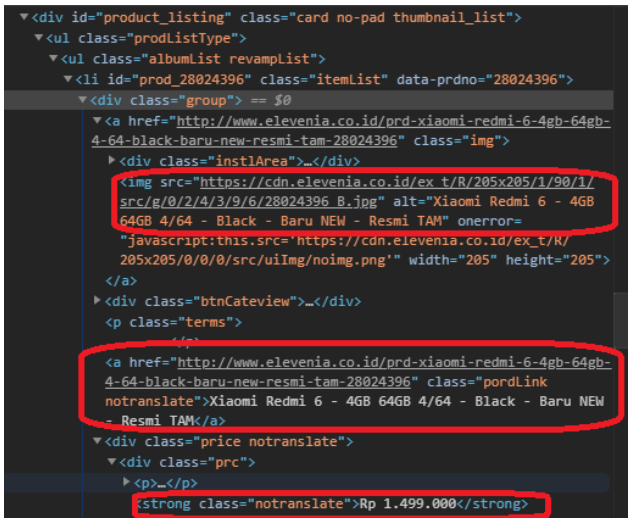
P-ISSN 2477-3891
E-ISSN 2615-4765

E. Proses *Web Scraping*

Proses yang dilakukan untuk elevenia, blanja, dan kaskus serupa sehingga akan dideskripsikan sekali pada bagian ini di situs Elevenia. Mula mula perlu diketahui DOM untuk ketiga situs dengan cara melakukan pencarian barang lalu gunakan fungsi *inspect elemen* pada website yang digunakan (biasanya firefox atau chrome). Lalu cari dan kerucutkan elemen terdekat dengan data yang akan disalin (*scrap*), kemudian cari elemen yang berkaitan dengan data yang bersangkutan beserta nama *class*.



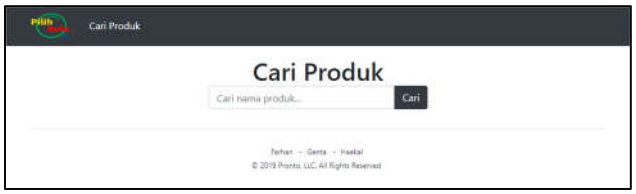
Gambar 6. Pemilihan elemen pada hasil pencarian



Gambar 7. Hasil Filtering DOM untuk Pengambilan Data

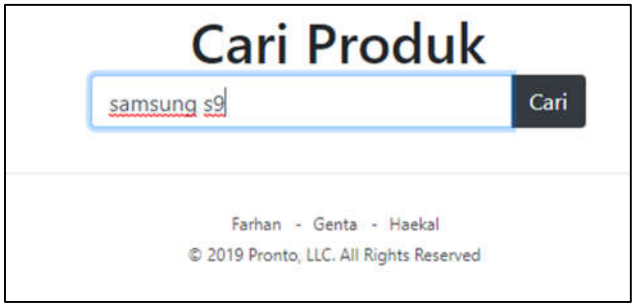
IV. HASIL DAN PEMBAHASAN

Aplikasi dibuat dengan implmentasi *web scraping* metode *Parsing DOM* dari situs *e-commerce*. Dibawah merupakan hasil dari aplikasi yang dibuat.



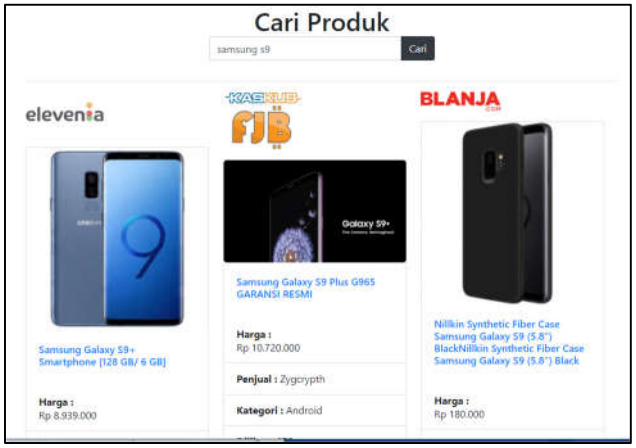
Gambar 8. Tampilan Awal Aplikasi PilihDulu

Di laman utama aplikasi PilihDulu yang telah kami buat, terdapat tempat untuk input text dengan tulisan cari nama produk, di kanannya terdapat tombol untuk mencari barang yang akan digunakan untuk kata kunci scraping pada 3 situs komersial online (*e-commerce*) diantaranya elevenia, forum jual beli kaskus, dan blanja.

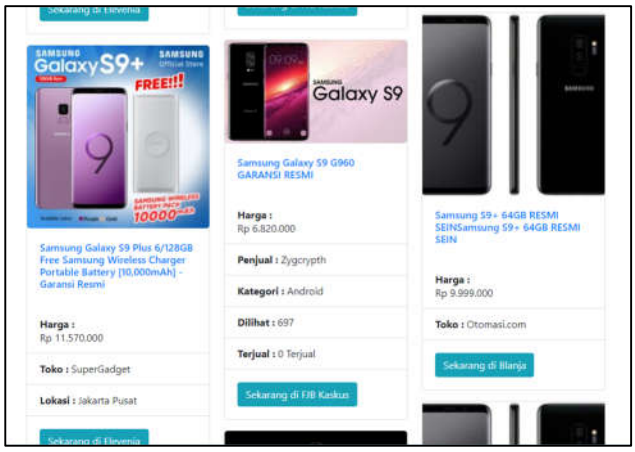


Gambar 9 Coba Pencarian

Dipilih kata kunci “samsung s9” sebagai *keyword* untuk scraping di tiga situs ini.



Gambar 10 Hasil Pencarian



Gambar 11. Hasil Pencarian

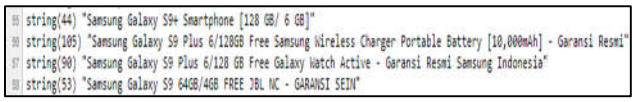
Setelah proses scraping selesai, aplikasi akan memunculkan semua data yang sesuai dengan *keyword* pada input sebelumnya. Untuk data foto, nama barang, harga, toko (untuk elevenia dan blanja), nama toko, lokasi (hanya Elevenia), dilihat dan terjualnya (hanya forum jual beli kaskus) merupakan hasil filtering *scraping* pada ketiga situs tersebut yang dilakukan secara terpisah. Dengan kata lain proses *web scraping* dilakukan pada salah satu situs terlebih dahulu hingga selesai baru kemudian dilakukan proses *web scraping* di situs selanjutnya hingga didapat data yang diinginkan dari ketiga situs tersebut.

Gambar 12 Isi dari pengambilan data HTML DOM

Pada gambar 8 ini merupakan hasil dari *dump* informasi variable yang menyimpan semua data DOM.

Gambar 13 Hasil Pengerucutan Berdasarkan Struktur

Semua data DOM tadi difilter / dikerucutkan berdasarkan struktur website *e*-commerce yang mengandung data barang didalamnya, misal pada suatu website terdapat tabel A terdapat rangkaian menu lalu pada tabel B terdapat list barang yang sedang laku pada website tersebut, untuk melakukan scraping berdasarkan DOM HTML, diambil deklarasi elemen dan nama kelas grup data yang ditargetkan contohnya nama barang dalam format ASCII pada tabel B maka elemen dan nama kelas tabel B pada HTML yang dipakai.



Gambar 14 Hasil Penyaringan Data, Melihat Nama Barang

Untuk pengambilan data spesifik seperti nama lengkap, harga, dan nama toko diperlukan filtering yang lebih spesifik lagi dengan penambahan elemen dan nama kelas hingga mencapai sumber data yang ditargetkan.

V. SIMPULAN

Berdasarkan hasil analisis dan pengujian terhadap penelitian ini didapatkan hasil tujuan yang tercapai, dimana:

1. Pengambilan data dari lebih dari satu situs komersial online (*e-commerce*) menggunakan teknik *parsing DOM* berhasil dilakukan dan hasil data yang diambil dapat digunakan.
2. Penampilan data dari ketiga situs komersial online (*e-commerce*) ini berhasil dilakukan dan dapat disuguhkan kepada pengguna.

VI. REFERENSI

[1] J. Agency, "Medium - The Most Effective Web Scraping Methods," JetRuby, 27 Ferbruari

2018. [Online]. Available: https://expertise.jetruby.com/the-most-effective-web-scraping-methods-62e7e34ada69. [Accessed Desember 2019].

[2] M. S. Rosa Ariani Sukamto, in Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek, Bandung , Informatika, 2013.

[3] P. W. Geoff Boeing, "New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings," Journal of Planning Education and Research , 2016.

[4] D. C. Tim Lee Berners, "w3," W3C, Juni 1993. [Online]. Available: https://www.w3.org/MarkUp/draft-ietf-iiir-html-01.txt. [Accessed Desember 2019].

[5] B. H. Elizabeth Castro, in HTML5 & CSS3 Visual QuickStart Guide (7th Edition), Peachpit Press, 2011, p. 550.

[6] "w3Schools.com What is HTML DOM," W3C, [Online]. Available: https://www.w3schools.com/whatis/whatis_htmldom.asp. [Accessed Desember 2019].

[7] D. Team, "dewaweb - Web Scraping : Panduan dan Teknik-Tekniknya," Dewaweb, 16 November 2018. [Online]. Available: https://www.dewaweb.com/blog/web-scraping-panduan-dan-teknik-tekniknya/. [Accessed December 2019].

[8] S. W. Ambler, in The Object Primer: Agile Model Driven Development with UML 2, Cambridge University Press, 2004.

[9] "W3 - About HTML," W3C, [Online]. Available: http://www.w3.org/html/wg/drafts/html/master/introduction.html#html-vs-xhtml. [Accessed Desember 2019].

[10] "Shield Square - Scrapping Techniques," radware, 2019. [Online]. Available: https://www.shieldsquare.com/what-are-the-different-scraping-techniques/. [Accessed Desember 2019].

[11] A. R. Pratama, "CodePolitan - Belajar Unified Modeling Language (UML) - Pengenalan," CodePolitan, 21 Januari 2019. [Online]. Available: https://www.codepolitan.com/unified-modeling-language-uml. [Accessed Desember 2019].

[12] R. Foxer, "Javan Cipta Solusi - Teknik Dasar Web Scraping," 4 April 2017. [Online]. Available: https://blog.javan.co.id/teknik-dasar-web-scraping-aa7d7e223093. [Accessed Desember 2019].

[13] R. Gunawan, A. Rahmatulloh, I. Darmawan, and F. Firdaus, "Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath," 2019. DOI: 10.2991/icoiese-18.2019.50

[14] Rahmatulloh, A., & Gunawan, R. (2020). Web Scraping with HTML DOM Method for Data Collection of Scientific Articles from Google Scholar. Indonesian Journal of Information Systems, 2(2), 16. https://doi.org/10.24002/ijis.v2i2.3029