Rancang Bangun Web Scraping Pada Media Online Berita Nasional

Muhammad Ahkam Adli[*1], Listra Firgia[2]
[1,2]Jurusan Teknik Informatika; STMIK Pontianak. Jl. Merdeka No.372 Pontianak, 0561-735555
e-mail: [*1]muhammadaadly@gmail.com, [2]Listra @stmikpontianak.ac.id

Abstrak

Perkembangan teknologi informasi yang maju seperti sekarang ini membuat masyarakat semakin cepat dalam mengakses informasi maupun berita. Seiring dengan banyaknya penyedia layanan situs berita online membuat pembaca harus berpindah-pindah situs berita untuk melihat berita yang berbobot dan berkualitas. Teknik scraping adalah suatu teknik yang digunakan untuk mengambil, menganalisa dan memproses suatu data dari suatu sistem atau dokumen yang berbeda. Aplikasi yang dibangun memanfaatkan teknologi web scraping untuk mengambil data pada berbagai situs berita nasional. Metode perancangan yang digunakan adalah Prototype, alat pemodelan system adalah Unified Modeling Language (UML). Teknik pengumpulan data dengan observasi dengan melakukan pengamatan dan menganalisis website yang telah ada sebagai referensi yang tepat untuk penulisan dan studi dokumentasi dengan cara pengumpulan data yang berkaitan dengan penerapan web scraping pada website melalui jurnal-jurnal ilmiah, dan ebook yang ada di internet yang berkaitan dengan penelitian ini. Hasil dari peneltian ini memudahkan pengguna dalam membaca berita. Pengguna tidak harus melihat berita per situs, hanya dengan satu website dapat melihat semua berita dari berbagai sumber dan memudahkan pengguna mencegah berita hoax.

Kata kunci— *Web Scraping, Berita, Prototype, UML*

Abstract

The development of advanced information technology, as it is now, makes people faster in accessing information and news. Along with the number of online news site service providers, readers must move to news sites to see quality and quality news. Scraping technique is a technique used to retrieve, analyze and process a data from a different system or document. The application built utilizes web scraping technology to retrieve data on various national news sites. The design method used is Prototype, a system modeling tool is the Unified Modeling Language (UML). Data collection techniques with observations by observing and analyzing existing websites as appropriate references for writing and documentation studies by collecting data relating to the application of web scraping on the website through scientific journals, and e-books on the internet relating to this research. The results of this study make it easier for users to read the news. Users do not have to see news per site, only with one website can see all the news from various sources and make it easier for users to prevent hoax news.

Keywords— *Web Scraping, News, Prototype, UML*

1. PENDAHULUAN

Perkembangan teknologi informasi yang maju seperti sekarang ini membuat masyarakat semakin cepat dalam mengakses informasi maupun berita. Masyarakat selalu menyerap informasi

maupun berita yang di baca kemudian menjadi bahan acuan atau referensi dalam kehidupan sehari-hari. Saat ini informasi yang sering di konsumsi masyarakat indonesia ialah informasi mengenai isu atau berita nasional, hal tersebut berbanding lurus dengan berkembangnya penyedia situs media berita online yang semakin banyak.

Seiring dengan banyaknya penyedia layanan situs berita online membuat pembaca harus berpindah-pindah situs berita untuk melihat berita yang berbobot dan berkualitas. Selain harus membuka banyak situs berita, iklan-iklan dalam situs berita juga mengganggu. Pembaca menjadi tidak fokus membaca berita karena adanya iklan dalam situs berita. Biasanya pembaca harus menutup terlebih dahulu iklan dalam bentuk pop up pada situs berita. Hal tersebut membuat tersitanya waktu dan merepotkan untuk membaca berita. Saat ini dibutuhkan suatu website khusus untuk menggabungkan beberapa situs tersebut menjadi satu. Teknik web scraping merupakan salah satu solusi yang cocok digunakan untuk mengambil berita dari berbagai situs tersebut.

Teknik scraping adalah suatu teknik yang digunakan untuk mengambil, menganalisa dan memproses suatu data dari suatu sistem atau dokumen yang berbeda. Teknik scraping biasanya digunakan untuk mengambil data dari website yang bisa disebut web scraping, Web Scraping juga merupakan proses pengambilan sebuah dokumen semi-terstruktur dari Internet, umumnya berupa halaman-halaman web dalam bahasa markup seperti HTML atau XHTML, dan menganalisis dokumen tersebut untuk diambil data tertentu dari halaman tersebut untuk digunakan bagi kepentingan lain[1]. Web scraping berhubungan dengan pengindeksan web yang merupakan suatu teknik universal yang dipakai hampir semua search engine. Perbedaannya web scraping lebih berfokus pada transformasi dari suatu web yang tidak terstruktur, umumnya dalam format HTML menjadi suatu format data terstruktur yang dapat disimpan dan dianalisa pada database atau lembar kerja[2]. Manfaat dari web scraping ialah agar informasi yang diambil atau digunakan lebih terfokus sehingga memudahkan dalam melakukan pencarian sesuatu[3].

Aplikasi yang dibangun memanfaatkan teknologi web scraping untuk mengambil data pada berbagai situs berita nasional. Web scraping sering dikenal sebagai screen scraping . Web Scraping tidak dapat dimasukkan dalam bidang data mining karena data mining menyiratkan upaya untuk memahami pola semantik atau tren dari sejumlah besar data yang telah diperoleh. Aplikasi web scraping (juga disebut *intelligent, automated, or autonomous agents*) hanya fokus pada cara memperoleh data melalui pengambilan dan ekstraksi data dengan ukuran data yang bervariasi[4]. Teknik scraping dapat dilakukan dengan cara mempelajari dokumen HTML dari website berita yang akan diambil informasinya untuk di tag HTML tujuannya ialah untuk mengapit informasi yang diambil, setelah itu mempelajari teknik navigasi pada website berita yang akan diambil informasinya untuk ditirukan pada aplikasi web scraping yang akan dibuat, kemudian aplikasi web scraping akan mengotomatisasi informasi yang didapat dari website berita yang telah ditentukan, informasi yang didapat tersebut akan disimpan ke dalam tabel basis data[5].

Penelitian ini bertujuan untuk mengekspolarasi laman-laman yang menyajikan berita dengan topik sosial politik kemudian mengumpulkan semua informasi yang ada di laman tersebut secara otomatis dengan menggunakan teknologi web scraping.. Dengan adanya web scraping ini memudahkan pengguna dalam membaca berita. Pengguna tidak harus melihat berita per situs, hanya dengan satu website dapat melihat semua berita dari berbagai sumber dan memudahkan pengguna mencegah berita *hoax*.

2. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah metode penelitian *Research and Development* (R&D). Metode penelitian Research and Development merupakan hasil dari penelitian pengembangan, tidak mengembangkan suatu produk baru atau menyempurnakan produk yang telah ada melainkan untuk menemukan pengetahuan atau jawaban atas permasalahan

praktis. Metode penelitian dan pengembangan juga didefinisikan sebagai suatu metode penelitian yang digunakan untuk menghasilkan produk tertentu, dan menguji keefektifan produk tersebut[6].

Metode perancangan sistem menggunakan *Prototype. Prototype* didefinisikan satu versi dari sebuah sistem potensial yang memberikan ide bagi para pengembang dan calon pengguna, bagaimana sistem akan berfungsi dalam bentuk yang telah selesai. Proses pembuatan prototype ini disebut *prototyping*[7]. Teknik pengumpulan data yang digunakan penulis yaitu observasi langsung dan studi dokumentasi. *Prototype* bukanlah merupakan sesuatu yang lengkap, tetapi sesuatu yang harus di evaluasi dan dimodifikasi kembali. Segala perubahan dapat terjadi *prototype* dibuat untuk memenuhi kebutuhan pengguna dan pada saat yang sama memungkinkan pengembang untuk lebih memehami kebutuhan pengguna secara lebih baik. Berikut merupakan langkah-langkah atau tahapan dalam metode prototype[8]:

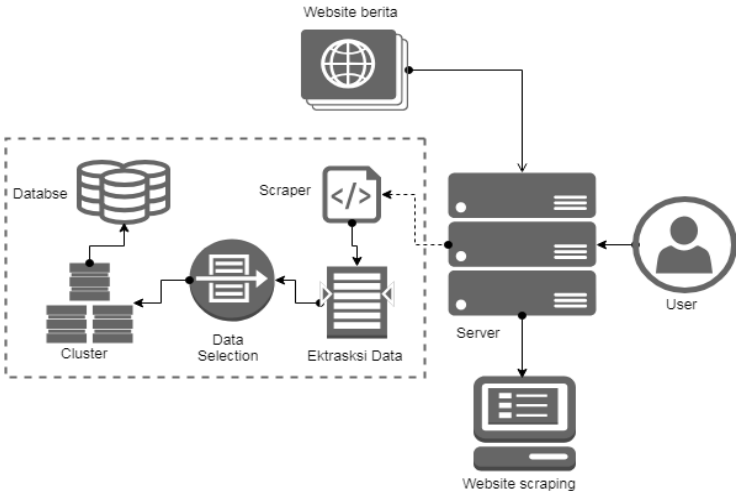
a. Komunikasi dan pengumpulan data awal, yaitu analisis terhadap kebutuhan pengguna.
b. Quick design, yaitu tahapan pembuatan design secara umum untuk selanjutnya dikembangkan kembali.
c. Pembentukan prototype, yaitu pembuatan perangkat prototype termasuk pengujian dan penyempurnaan.
d. Evaluasi terhadap prototype, yaitu mengevaluasi prototype dan memperhalus analisis terhadap kebutuhan pengguna.
e. Perbaikan prototype, yaitu pembuatan tipe yang sebenernya berdasarkan hasil dari evaluasi prototype.
f. Produksi akhir, yaitu memproduksi perangkat secara benar sehingga dapat digunakan oleh pengguna.

3. HASIL DAN PEMBAHASAN

Penelitian yang akan dilakukan dalam mengembangkan aplikasi web scraping ini menggunakan metode Prototype. Prototype merupakan proses pembuatan model sederhana software yang mengijinkan pengguna memiliki gambaran dasar tentang program serta melakukan pengujian awal. Prototype memberikan fasilitas bagi pengembang dan pemakai untuk saling berinteraksi selama proses pembuatan, sehingga pengembang dapat dengan mudah memodelkan perangkat lunak yang akan di buat. Pendekatan pengembangan ini memudahkan peneliti dalam merancang aplikasi. Metode prototype merupakan salah satu jenis metode pengembangan sistem yang sifatnya sangat cepat dan dapat menghemat waktu.

3.1 Model Arsitektur Sistem

Arsitektur dari sistem merupakan sekumpulan dari model-model terhubung yang menggambarkan sifat dasar dari sebuah sistem. Keanekaragaman dari banyak model menggambarkan bagian berbeda dan aspek atau pandangan yang berbeda dari suatu sistem. Perancangan model arsitektur web scraping mengindentifikasikan semua struktur sistem, prinsip komponen (sub-sistem/modul), hubungannya dan bagaimana didistribusikan.



Gambar 1. Arsitektur Sistem

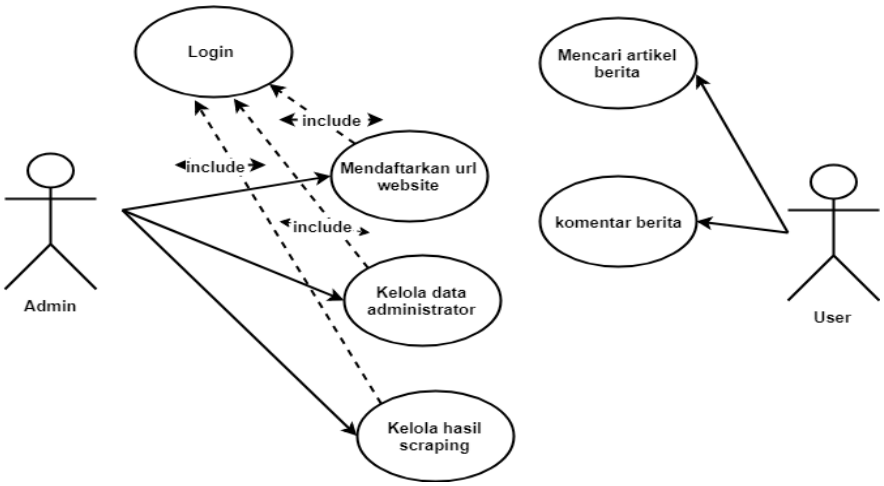
Gambar 1 menjelaskan desain sistem secara keseluruhan website scraping. *Server* melakukan *scraping* konten berita pada website detik.com, cnnindonesia.com dan kompas yang di scraping pada website dan mengambil apa yang dibutuhkan dan menghilangkan bagian yang tidak perlu, kemudian data di seleksi dan disimpan di database.

3.2 Model Arsitektur Web Scraping

Tahapan perancangan web scraping mengacu pada perancangan berbasis obyek. Startegi ini dalam istilah aslinya disebut sebagai OOD (*Object Orianted Design*) dan dianggap menjadi startegi perancanaan paling modern. Dalam penelitian ini penulis menggunakan UML (*Unified Modeling Language*). Berikut ini adalah perancangan astitektur perangkat lunak yang dimodelkan dengan UML :

3.2.1 Use Case Diagram

Use Case Diagram digunakan untuk menentukan kebutuhan apa saja yang diperlukan dari suatu sistem. Jadi, dapat digambarkan dengan rinci bagaimana suatu sistem memproses atau melakukan sesuatu, bagaimana cara actor akan menggunakan sistem, serta apa saja yang dapat dilakukan terhadap suatu sistem.

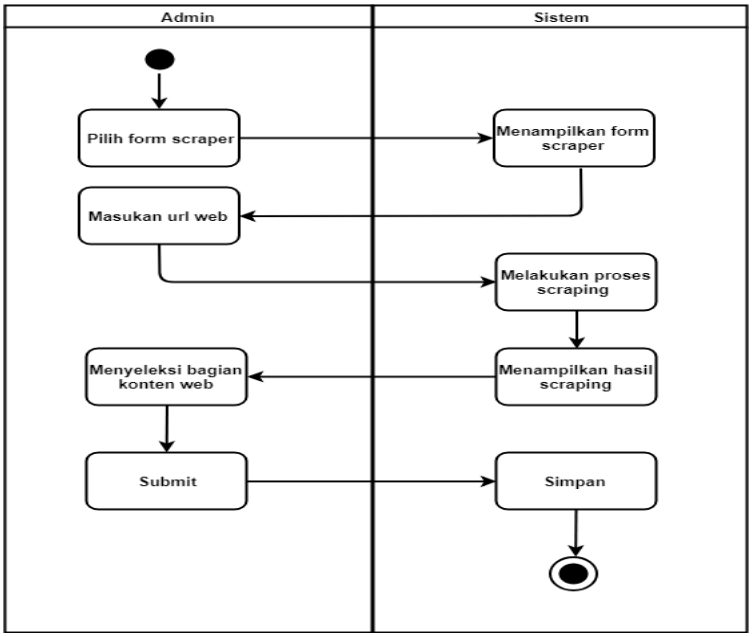


Gambar 2. Use Case Diagram Website Scraping

Use case diagram website scraping terdiri dari admin dan user. Actor admin bertugas bertugas untuk memanajemen isi dari website secara keseluruhan seperti mendaftarkan url website berita yang ingin di scraping, mengelola data admin dan mengelola hasil scraping. Actor user hanya diperbolehkan mencari artikel berita dan memberikan komentar.

3.2.2 Activity Diagram

Activity Diagram menggambarkan berbagai alur aktivitas dalam sistem yang sedang dirancang, bagaimana masing-masing alur berawal, decision yang mungkin terjadi, dan bagaimana mereka berakhir. Activity Diagram juga dapat menggambarkan proses paralel yang mungkin terjadi pada beberapa eksekusi. Activity Diagram merupakan state diagram khusus, dimana sebagian besar state adalah action dan sebagian besar transisi di-trigger oleh selesainya state sebelumnya (*internal processing*). Oleh karena itu Activity Diagram tidak menggambarkan behaviour internal sebuah sistem (dan interaksi antar subsistem) secara eksak, tetapi lebih menggambarkan proses-proses dan jalur-jalur aktivitas dari level atas secara umum.



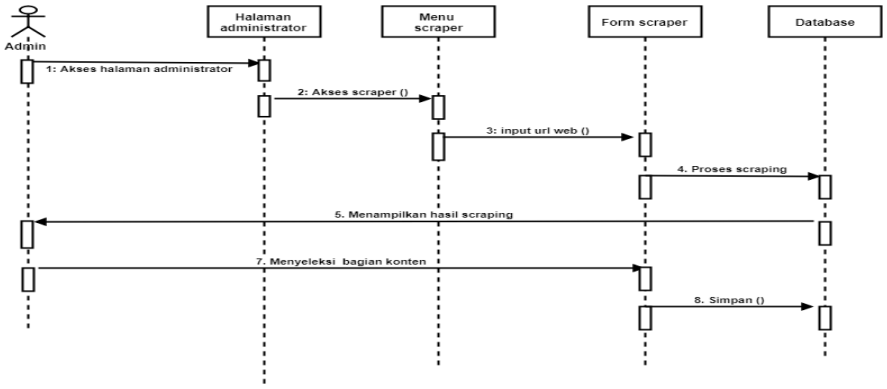
Gambar 3. Activity Diagram Mendaftarkan Url Website

Activity diagram mendaftarkan url websitte merupakan aktivitas dimana admin yang telah login dan masuk pada halaman administrator. Activity dimulai dari pemilihan form scraper oleh admin. Pada form scraper, admin dapat memasukan url web berita yang di scraping, sistem akan melakukan proses scraping. Sistem akan menampilkan hasil dari proses scraping dan admin dapat menyeleksi bagian konten website yang akan di tampilkan. Kemudian admin meklik tombol simpan dan konten akan tampil pada bagian halaman website.

3.2.3 Sequence Diagram

Sequence diagram menggambarkan interaksi antar objek di dalam dan di sekitar sistem (termasuk pengguna, display, dan sebagainya) berupa message yang digambarkan terhadap waktu. Sequence diagram terdiri atar dimensi vertikal (waktu) dan dimensi horizontal (objek-objek yang terkait). Sequence diagram dapat digunakan untuk menggambarkan skenario atau rangkaian langkah-langkah yang dilakukan sebagai respons dari sebuah event untuk

menghasilkan output tertentu. Diawali dari apa yang men-trigger aktivitas tersebut, proses dan perubahan apa saja yang terjadi secara internal dan output apa yang dihasilkan. Masing-masing objek, termasuk aktor, memiliki lifeline vertikal. Message digambarkan sebagai garis berpanah dari satu objek ke objek lainnya. Pada fase desain berikutnya, message akan dipetakan menjadi operasi/metoda dari class.

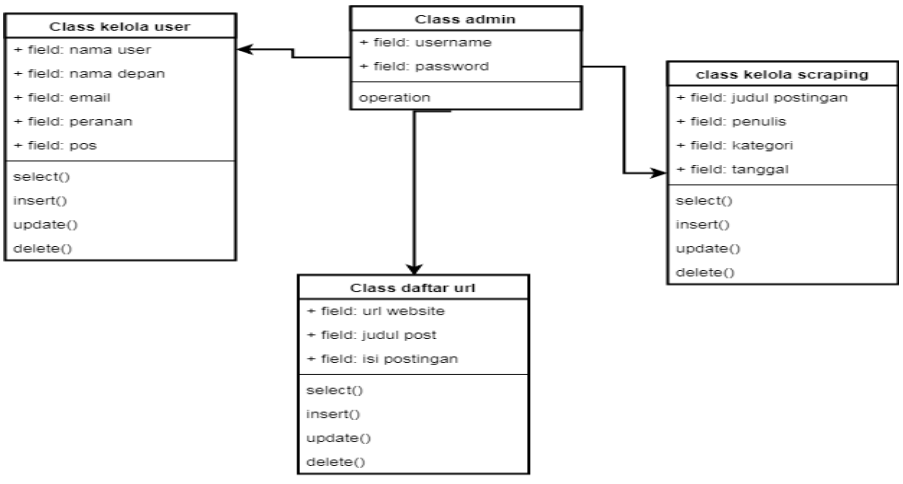


Gambar 4. Sequence Diagram Mendaftarkan Url Website

Sequence diagram mendaftarkan url merupakan interaksi antara admin dengan sistem scraping. Pada form scraper, admin mengisikan url website dan sistem akan melakukan proses scraping. Sistem akan menampilkan hasil scraping website dan admin dapat melakukan seleksi bagian konten yang akan di tampilkan pada halaman utama website scraping.

3.2.4 Class Diagram

Class diagram adalah diagram yang digunakan untuk menampilkan beberapa kelas serta paket-paket yang ada dalam sistem atau perangkat lunak yang sedang kita gunakan. Class diagram juga memberikan gambaran (diagram statis) tentang system atau perangkat lunak dan relasi-relasi yang ada didalamnya.



Gambar 5. Class Diagram Web scraping

Class diagram Website scraping barang menggambarkan hubungan antara entitas yang terkait dengan sistem penjualan barang. Pada sistem ini terdapat 4 entitas yang saling berelasi antara satu dengan yang lainnya. Relasi ini menggambarkan bahwa ada kaitan secara langsung maupun tidak langsung diantara setiap entitas sistem.

3.3 Perancangan Database

Merancang struktur tabel database berguna memenuhi kebutuhan saat ini dan kemudahannya untuk dikembangkan pada masa yang akan datang. Perancangan model konseptual, digunakan konsep pendekatan rasional namun tidak berarti konsep ini harus diimplementasikan kemodel relasional saja tetapi juga dapat dengan model hirarki dan network model. Model konseptual tidak tergantung aplikasi tertentu dan tidak tergantung pada DBMS dan hardware yang digunakan. Pada perancangan model konseptual tinjauan dilakukan pada struktur data dan relasi antar file menggunakan model dan relasional.

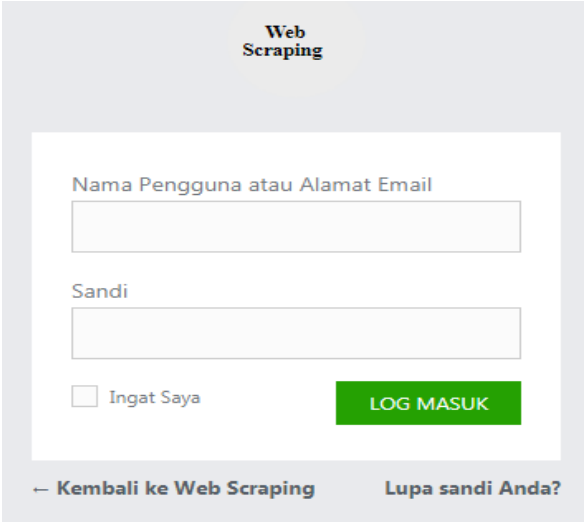
Tabel 1. Spesifikasi Tabel Posttingan

Collumn	Type	Null	Default	Index
ID	bigint(20)	No		Primary key
post_author	bigint(20)	No		
post_date	datetime	No	0000-00-00	
post_content	longtext	No		
post_title	text	No		
post_excerpt	text	No		
post_status	varchar(20)	No		
comment_status	varchar(20)	No		
post_name	varchar(255)	No		
post_type	varchar(20)	No		
post_modified	datetime	No	0000-00-00	
post_coontent_filtered	datetime	No	0000-00-00	

3.4 Interface Desain Website Scraping

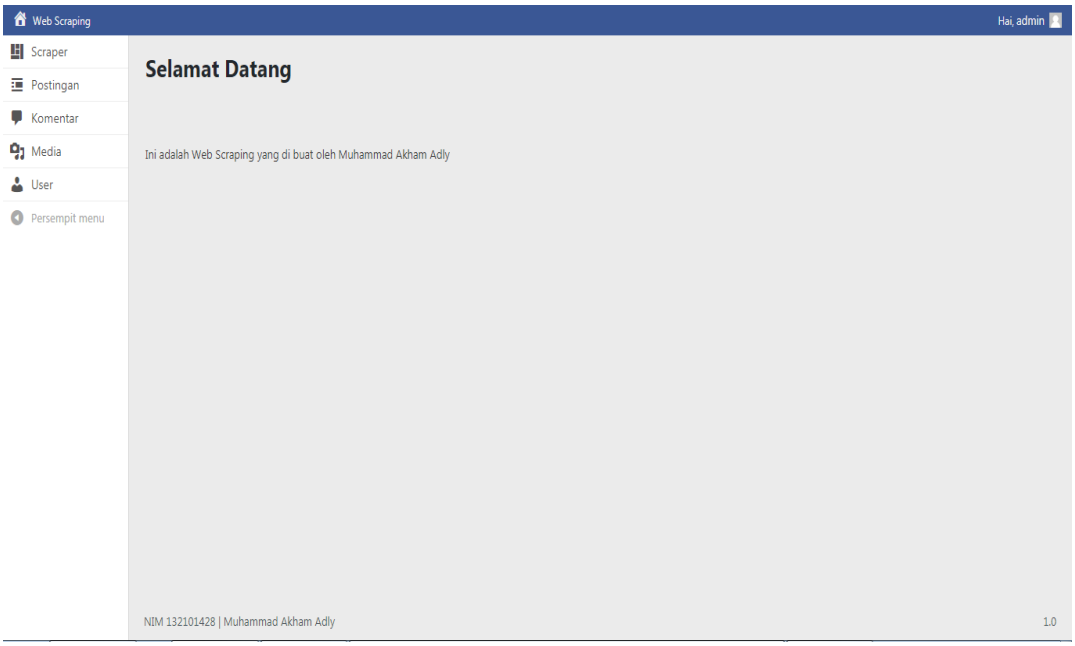
Merancang antarmuka merupakan bagian yang paling penting dari merancang sistem. Biasanya hal tersebut juga merupakan bagian yang paling sulit karena dalam merancang antarmuka harus memenuhi tiga persyaratan: sebuah antarmuka harus sederhana, sebuah antarmuka harus lengkap, dan sebuah antarmuka harus memilki kinerja yang cepat. Alasan utama mengapa antarmuka sulit untuk dirancang adalah karena setiap antarmuka adalah sebuah bahasa pemrograman yang kecil: antarmuka menjelaskan sekumpulan objek-objek dan operasi-operasi yang bisa digunakan untuk memanipulasi objek.

Kontruksi form login admin adalah bentuk otentikasi user login ke web. Dengan form login seorang administrator dapat menggunakan fasilitas khusus yang disediakan oleh sistem untuk melakukan manipulasi data seperti penambahan data, perubahan data, pencarian data dan penghapusan data. Berikut ini adalah rancangan form login admin:



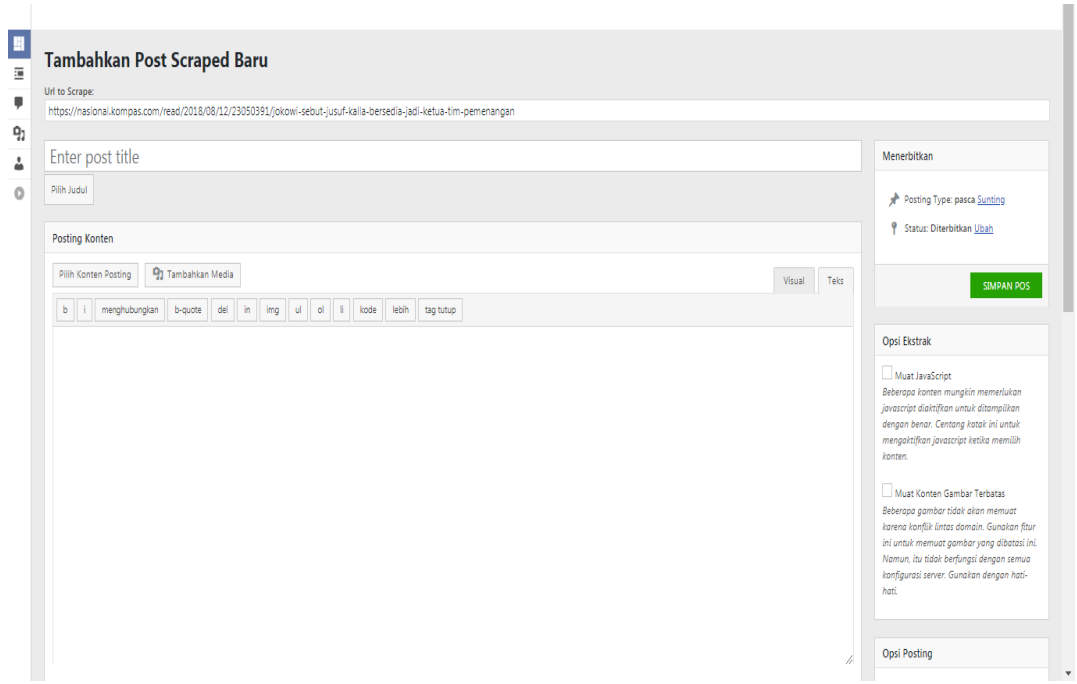
Gambar 6. Form Login Admin

Konstruksi form menu admin didesain sebagai tempat utama untuk adminisrator web melakukan kegiatan pengelolaan website seperti pengelolaan data web scraping dan data administator. Berikut ini adalah desain form menu utama admin:



Gambar 7. Kontruksi form menu admin

Konstruksi form halaman konfirmasi tambah post scraping didesain untuk menambahkan artikel berita baru dengan mengisikan url website yang ingin di scraping, kemudian sistem otomatis menampilkan hasil scraping. Berikut ini adalah form halaman form tambah post scraping :

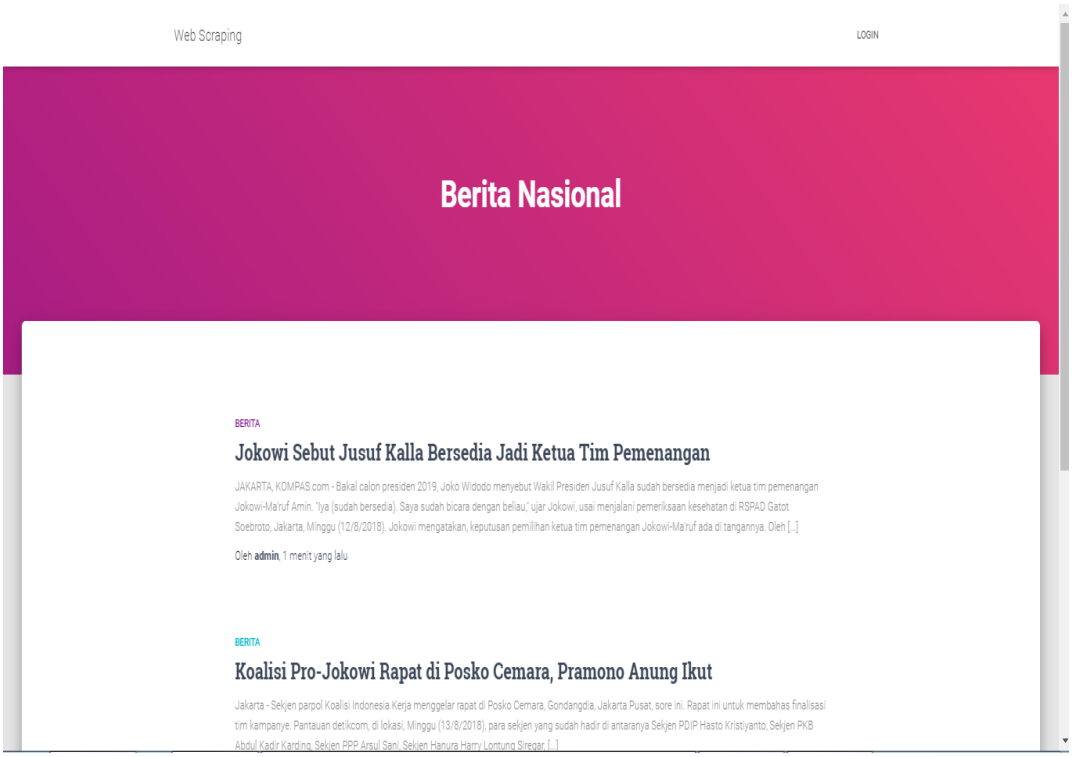


Gambar 8. Kontruksi form tambah post scraping

Konstruksi form seleksi konten didesain oleh admin untuk menyeleksi bagian konten dari website yang telah di scraping. Admin dapat memilih bagian dari konten yang perlu di ambil untuk di tampilkan ke website scraping. Berikut ini adalah form halaman form tambah post scraping :

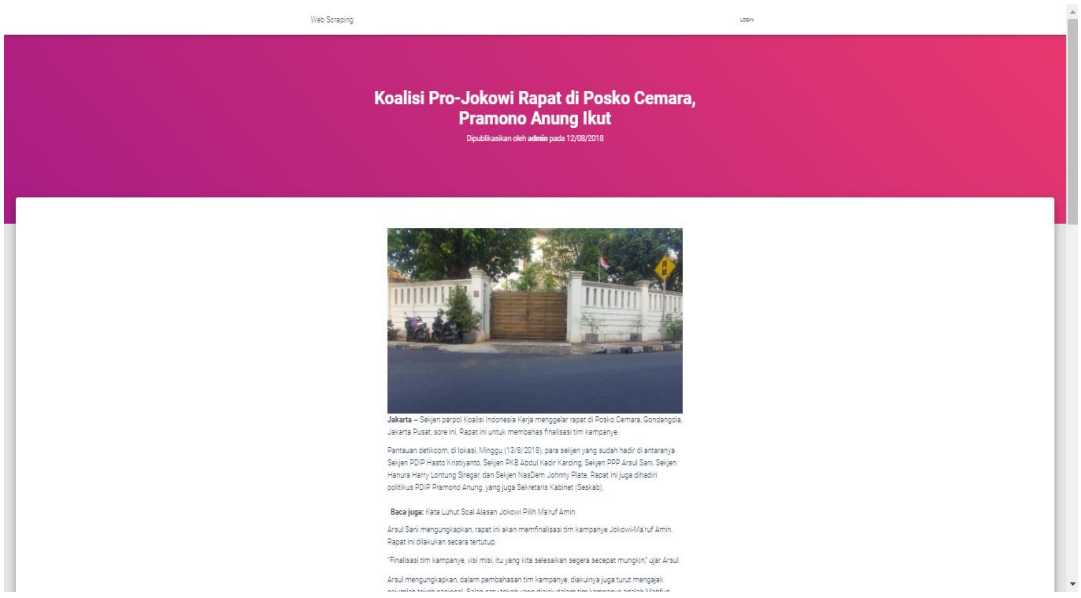
Gambar 9. Rancangan form seleksi konten

Konstruksi halaman depan website didesain untuk menampilkan berita hasil dari proses scraping pada website berita nasional. Para pengunjung dapat melihat berita yang telah discraping. Di kanan atas terdapat fitur login yang hanya bisa diakses oleh pengguna yg terdaftar sebagai admin. Berikut ini adalah form halaman depan web scraping :



Gambar 10. Rancangan form halaman utama

Halaman detil berita didesain untuk menampilkan berita secara detil dan pada halaman ini juga tersebut tombol untuk melakukan komentar dan membagikan berita. Berikut ini adalah halaman detil berita:



Gambar 11. Rancangan form detil berita

4. KESIMPULAN

Berdasarkan hasil kajian dan pembahasan dalam penelitian ini, maka kesimpulan dari penelitian ini adalah penelitian ini dilakukan atas dasar tujuan yaitu menghasilkan website

scraping yang memudahkan pengguna dalam membaca berita hanya pada satu situs. Pemodelan perangkat lunak digambarkan dengan model UML yang terdiri dari use case diagram, activity diagram, sequence diagram dan class diagram. Website berhasil menyimpan otomatis data berita hasil scraping pada database. Memperoleh data berita yang terstruktur dan mengumpulkan data pada situs berita nasional agar informasi yang diambil lebih terfokus. Hasil akhir website scraping adalah sebuah website yang dapat menggambil konten dari beberapa situs website berita kemudian digabungkan kedalam satu website.

5. SARAN

Berdasarkan hasil pembahasan dan kesimpulan, maka saran dari penelitian ini adalah Perlu dilakukan pengembangan lebih lanjut guna mencapai hasil yang maksimal. Penelitian lanjut bisa didasarkan pada penerapan metode yang lebih lengkap lagi agar bisa mencapai tujuan yang maksimal dan flesibilitas perlu ditingkatkan lagi agar memberikan kemudahan bagi pengunjung website yang menggunakan perangkat mobile, harus mengedepankan interaktif agar memberikan kenyamanan bagi pengunjung dalam berinteraksi dengan web kemudian Perlu adanya fitur yang dapat menscraping berita secara otomatis.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada kedua orang tua beserta keluarga besar, dan seluruh sahabat seperjuangan angkatan 2013 STMIK Pontianak yang telah memberikan dukungan baik moral maupun materil kepada penulis untuk dapat menyelesaikan penelitian ini dan penulis juga mengucapkan terima kasih kepada reviwer atas bimbingan dan arahan sehingga tulisan ini dapat sesuai dengan seperti apa yang diharapkan.

DAFTAR PUSTAKA

[1] Turland, Matthew., 2010, *Guide to Web Scraping with PHP*. Marco Tabini & Associates.Inc, Canada.

[2] Mitra, Vivensius, Sujaini, Herry, Negara, A. B. P., Rancang Bangun Aplikasi Web Scraping Untuk Korpus Paralel Indonesia - Inggris Dengan Metode HTML DOM, *Jurnal Sistem dan Teknologi Informasi (JUSTIN,)* Volume 1, No 1, 2017.

[3] Batubara, M. S., Implementasi Ekstraksi Web (Web Scraping) Pada Mesin Pecari Jurnal Ilmiah Menggunakan Metode Ekspresi Reguler, *JITEKH (Jurnal Ilmiah Teknologi Harapan)*, Volume 1, Maret 2016.

[4] Josi, Ahmad, Abdillah, L. A., Suryayusra, Penerapan Teknik Web Scraping Pada Mesin Pencarl Artikel Ilmiah, *Jurnal Sisfo*, Volume 5, No 2, September 2014.

[5] Juliasari, Noni, Sitompul, J. C., Aplikasi Search Engine dengan Metode Depth First Search (DFS), *Jurnal BIT*, Volume (9):10, April 2012.

[6] Sugiyono, 2008, *Metode Penelitian Kunatitatif Kualitatif dan R&D*, Alfabeta, Bandung

[7] McLeod Jr., Raymond. 2013, *Sistem Informasi Manajemen*, Edisi Ketujuh, PT Prenhallindo, Jakarta.

[8] Hafizd, K. H., dan Sayyidati, Rabini, Sistem Informasi Perpustakaan Politeknik Negeri Tanah Laut, *Jurnal Sains dan Informatika*, Vol 3, November 2017.