PERANCANGAN APLIKASI WEB SCRAPING UNTUK KOLEKSI KONTEN RESEP MASAKAN TRADISIONAL JAWA BERBASIS XML

Setyawan Wibisono [1)], Mardi Siswo Utomo [2)]
Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Stikubank Semarang,
Jl. Tri Lomba Juang No. 1 Semarang, 50244
E-mail: [1)] setya.sonny@gmail.com, [2)] mardiutomo@gmail.com

Abstrak

Proses untuk memisahkan konten utama halaman situs dengan bagian-bagian yang tidak berhubungan dengan isi disebut dengan *scraping*. Dengan teknik ini konten utama dari suatu halaman situs dapat diekstrak, dikoleksi dan selanjutnya dapat diproses oleh proses pengindekan. Sistem ini adalah perangkat lunak berbasis web dengan tujuan melakukan pengambilan isi dari konten halaman web. Hal-hal yang dapat diwujudkan dalam sistem ini diantaranya: (1) Sistem dapat secara otomatis mengekstrak konten utama dari suatu halaman web, (2) Dalam penelitian ini digunakan halaman dokumen pada situs resmi sebuah produk makanan dengan merk Bango, (3) Pengambilan data/*crawling Uniform Resource Locator* (URL) pada situs resmi sebuah produk makanan merk Bango menggunakan aplikasi sphider, (4) Hasil *scraping* resep disimpan dalam basisdata, (5) Sistem ini dapat memproduksi data resep dengan format XML (*eXtensible Markup Language*), (6) Aplikasi diintegrasikan dalam bentuk *plugin* CMS wordpress yang dapat diunduh di secara bebas (7) Sistem diimplementasikan secara *online* menggunakan sebuah situs yang telah disiapkan. Teknik *web scraping* dapat digunakan untuk mengambil konten resep pada situs pada berbagai situs yang memuat resep masakan. Penyimpanan resep ke dalam basisdata, mempermudah transformasi data ke bentuk lainnya.

Kata kunci— konten, resep, scraping, XML

Pendahuluan

Di internet banyak sekali data tentang informasi dan pengetahuan yang dapat dengan mudah didapat, tetapi data semacam ini begitu heterogen bentuk dan formatnya sehingga sangat sulit untuk dianalisa secara langsung. Data dari internet biasanya berupa halaman situs menggunakan format HTML, dengan sebagian isi halaman tersebut merupakan informasi untuk pengguna manusia seperti tombol navigasi, pencarian gambar dan layout untuk memperindah dan mempermudah situs untuk dibaca. Bagian ini biasanya tidak dibutuhkan dalam proses analisa temu kembali informasi bahkan bisa dianggap sebagai *noise* karena bisa menurunkan kualitas hasil.

Konten utama halaman situs biasanya terletak di tengah halaman, dimana bagian ini sudah umum diasumsikan sebagai intisari dari halaman tersebut oleh pengguna. Deteksi konten utama merupakan hal terpenting dalam proses ekstraksi konten utama untuk dipisahkan dari bagian lain seperti *header*, *footer* dan *sidebar*. Konten bukan data teks dieliminasi dalam proses ekstraksi konten utama, sehingga didapatkan data yang akurat sesuai dengan maksud dari halama situs tersebut. Data semacam ini dapat digunakan sebagai korpus pada uji coba sistem temu kembali. Proses untuk memisahkan konten utama halaman situs dengan bagian yang tidak berhubungan dengan isi disebut dengan *scraping*. Dengan teknik ini konten utama suatu halaman situs dapat diekstrak, dikoleksi dan selanjutnya dapat diproses oleh proses pengindekan.

Teknik *scraping* dapat dilakukan dengan diantaranya menggunakan analisa HTML DOM (*document object model)* dan dengan menggunakan teknik pemrograman regular ekspresi. Kedua teknik ini mempunyai keunggulan tersendiri dan memberikan hasil yang tidak jauh berbeda. Pada teknik DOM dibutuhkan *XQuery* untuk mengekstrak konten utama dari halaman situs sedangkan pada teknik regular ekspresi menggunakan metoda penentuan pola yang menjadi awal dan akhir suatu konten utama pada halaman situs. Hasil dari proses *scraping* dapat disimpan dalam berbagai macam format, dimana dalam penelitian ini hasil proses *scraping* akan disimpan dengan format XML. Diharapkan dengan data berformat XML ini, data akan lebih mudah untuk digunakan sebagai korpus pada temu kembali informasi (Utomo, 2013).

Studi Pustaka

Bing (2011) menyatakan bahwa w*eb mining* (pertambangan web) seringkali disebut *web extraction* ataupun *web scraping* bertujuan untuk menemukan informasi yang berguna atau

pengetahuan dari struktur *hyperlink web*, konten halaman, dan penggunaan data. Berdasarkan jenis utama dari data yang digunakan dalam proses pertambangan web, maka pertambangan web dapat dikategorikan menjadi tiga jenis: pertambangan struktur web, pertambangan konten web dan pertambangan penggunaan web. *Web data Extraction* (ekstraksi data web) adalah proses penggalian informasi terstruktur dari sumber data web terstruktur atau semi terstruktur. Ekstraksi web juga disebut sebagai *web data mining* atau *web scraping*. Chaudhari dan Paikrao (2012) merancang perangkat lunak dengan fungsi mengekstrak halaman web pra target yang berisi data yang diinginkan dengan bantuan robot dan *crawler web* yang memberikan petunjuk tentang apa yang harus dicari dan untuk apa. Ketika mencari halaman dari sebuah situs web, perangkat lunak juga akan mengikuti link apapun yang dapat menghubungkan dengan konten lain yang relevan.

Proses untuk menemukan informasi yang berguna dapat dilakukan dengan metoda *data scraping* yang mencakup sejumlah metoda yang berbeda untuk mendapatkan data dari situs web atau basisdata yang biasanya dilakukan dengan menggunakan perangkat lunak. Jennings dan Yates, (2008) menyatakan salah satu metoda yang digunakan adalah *screen scraping*, yaitu program *scraper* hanya melakukan ekstraksi data kunci yang muncul pada tampilan layar. Program *screen scraping* akan mengabaikan bagian *coding* dan hanya mencari dan melakukan ekstraksi *plain text* dari sebuah halaman web. Disebut juga *"web harvesting"*, yang melibatkan penggunaan program *scraper* untuk mengekstraksi semua data yang berhubungan dengan struktur yang mendasari skrip HTML.

Dalam banyak penerapan, *web scraping* berguna untuk mendapatkan data dalam bentuk teks dari situs web lain dalam jumlah yang relatif besar. Dalam kaitannya dengan format data teks, maka format data XML dapat digunakan sebagai konten data yang diolah dan ditampilkan kembali dalam situs web yang menggunakan data hasil *scrape* dari situs web lain. W3C.org menyatakan bahwa XML adalah bahasa *markup* untuk dokumen yang berisi informasi yang terstruktur. Informasi yang terstruktur berisi kata-kata, gambar dan beberapa indikasi peran apa yang dimainkan konten, misalnya isi di bagian judul memiliki arti yang berbeda dari konten dalam sebuah catatan kaki, yang berarti sesuatu yang berbeda dari konten angka atau konten sebuah tabel basisdata. Sebuah bahasa *markup* adalah mekanisme untuk mengidentifikasi struktur dalam sebuah dokumen. XML mendefinisikan cara standar untuk menambahkan *markup* ke dokumen

Metodologi Penelitian

Metoda penelitian perancangan aplikasi web scraping untuk koleksi konten resep masakan tradisional Jawa berbasis XML menggunakan model prototyping(Pressman, 1997). Tahapan pengembangan sistem terdiri dari

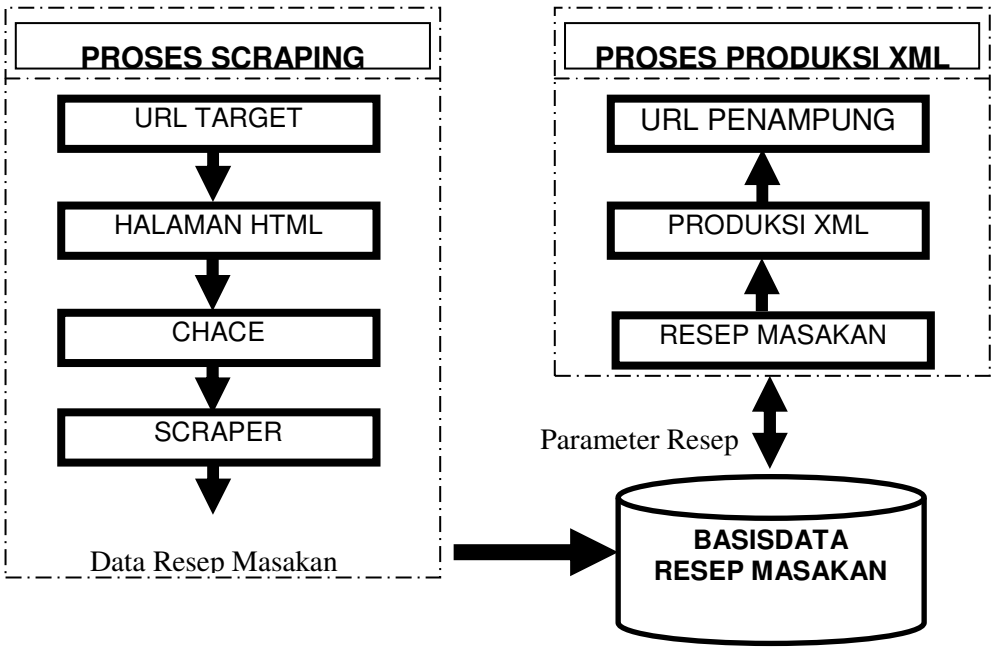
1. Analisis, pada saat ini di Indonesia, dokumentasi warisan budaya resep masakan tradisional Jawa dalam format XML belum ada. Dengan demikikan dibutuhkan sebuah sistem yang mampu mengumpulkan data resep masakan tradisional Jawa secara otomatis menggunakan perangkat lunak. Hasil pengumpulan data resep dapat digunakan sebagai salah satu sumber dalam pelestarian budaya takbenda. pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya.
2. Desain, pada tahap ini dibangun rancangan perangkat lunak komputer berbasis web yang bertujuan untuk melakukan pengambilan isi dari konten halaman web. Hal-hal yang diharapkan oleh pengguna agar dapat diwujudkan dalam sistem ini di antaranya adalah: a) Sistem dapat secara otomatis mengekstrak konten utama dari suatu halaman web, dalam penelitian ini digunakan halaman dokumen pada situs http://bango.co.id; b)Pengambilan data (*crawling*) URL pada http://bango.co.id menggunakan aplikasi sphider; c) Hasil *scraping* resep disimpan dalam basisdata; d) Sistem ini dapat memproduksi data resep dengan format XML; e) Aplikasi diintegrasikan dalam bentuk *plugin* CMS wordpress yang dapat diunduh di http://wordpress.org; f) Sistem diimplementasikan secara *online* menggunakan URL http://masakbagus.com.
3. Prototype, pada tahap ini dibangun aplikasi web scraping untuk koleksi konten resep masakan tradisional Jawa berbasis XML sesuai dengan disain dan kebutuhan sistem.

4. Pengujian, pada tahap ini dilakukan pengujian pada aplikasi yang sudah dibangun, pengujian menggunakan input query dalam bentuk teks dan kesesuaian query dengan hasil tampilan dan hasil dokumen yang di dapat.
5. Evaluasi, pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian – penyesuaian sesuai kebutuhan.
6. Penyesuaian : Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan.

Arsitektur sistem

Secara garis besar sistem ini terdiri dari dua bagian utama, yaitu sistem yang dapat mengekstrak data resep halaman situs http://bango.co/id kemudian menyimpannya dalam basis data, serta sistem yang dapat memproduksi data resep dengan format XML. Pada gambar 1 diperlihatkan halaman web diambil dari URL target yaitu http://bango.co.id, kemudian akan dibaca halaman HTML dan disimpan di *chace*. Bagian *scraper* mengkestrak resep masakan Jawa yang terdapat pada halaman tersebut dan menyimpannya ke dalam basis data.

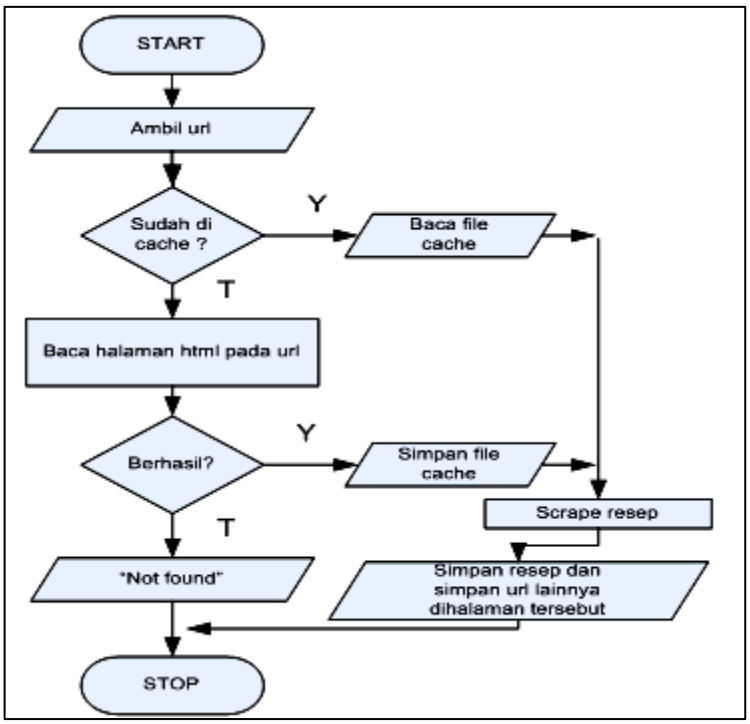
Ketika akan memproduksi resep dengan format XML, maka data resep diambil dari basisdata. Resep makanan yang diperoleh akan dibentuk menjadi format XML dan dapat ditampilkan pada URL penampung, dalam hal ini http://masakbagus.com. Aplikasi ditanam pada *web server* yang terkoneksi dengan jaringan internet. Aplikasi berjalan menggunakan *service http* dengan format transaksi data html, sehingga dapat dibuka menggunakan terminal yang terkoneksi ke jaringan komputer dan mampu/mempunyai *browser* WEB. Pengguna dapat melihat dokumen yang telah diekstrak dalam format XML.



Gambar 1. Arsitektur web scraping untuk koleksi konten resep masakan tradisional jawa berbasis XML

Diagram Alir Proses *Scraping*

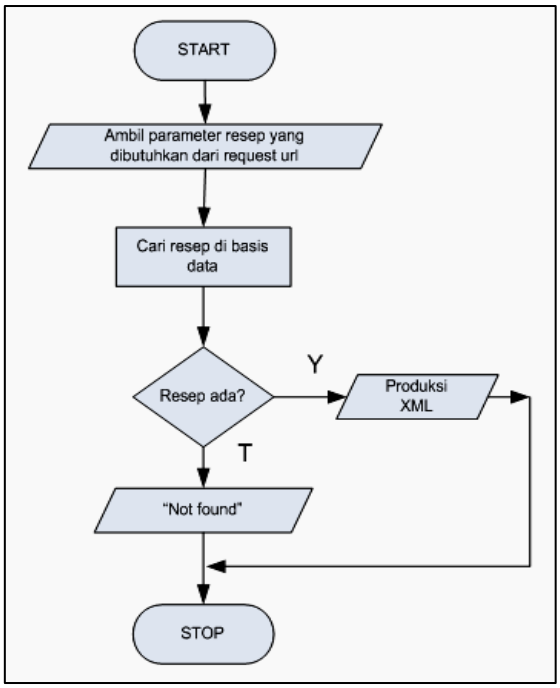
Pada gambar 2, aliran proses *scraping* dimulai dari pengambilan URL target kemudian mengambil konten dari URL tersebut. Setelah konten berhasil diambil terlebih dahulu disimpan ke file dengan nama file hasil *hashing* URL tersebut untuk keperluan *caching*. Kemudian konten resep diekstrak pada halaman situs URL tersebut. Hasil dari ekstrak langsung disimpan pada basisdata di masing-masing tabel yang tersedia. Pemilihan URL target masih dilakukan secara manual untuk menentukan bahwa halaman tersebut merupakan masakan Jawa atau tidak.



Gambar 2. Diagram aliran proses *scraping*

Diagram Alir Proses Produksi XML

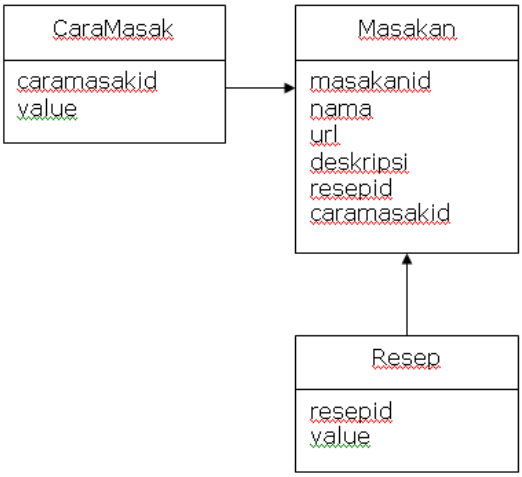
Pada gambar 3 diperlihatkan proses produksi data XML, diawali dengan membaca parameter URL produksi untuk menentukan resep mana yang akan diproduksi format XML-nya. Parameter ditentukan oleh pengguna sesuai kebutuhan. Setelah diketahui resep mana yang akan diambil, sistem akan mengambil data dari basisdata dan dengan bantuan *script* dapat merubah format dari basisdata menjadi data berformat XML.



Gambar 3. Diagram aliran proses produksi data XML

Rancangan Basisdata

Rancangan basisdata aplikasi web scraping untuk koleksi konten resep masakan tradisional Jawa berbasis XML diperlihatkan pada gambar 4.



Gambar 4. Rancangan basisdata

Hasil dan Pembahasan

Fungsi *Web Scraping*

Diagram alir pada gambar 2 diperlihatkan alur proses dari fungsi *web scraping*, dimana fungsi terlebih dahulu melakukan pengambilan alamat URL dari basis data yang telah disiapkan. Proses pengambilan halaman atau biasa disebut dengan *fetching* atau pengunduhan dapat dilakukan dengan perintah PHP *file_get_content*. Metode lainnya untuk mengambil konten web adalah CURL, metode ini dapat mengirim informasi lengkap dan detail layaknya sebuah *web browser* sehingga *web server* menganggap permintaan dilakukan oleh seorang pengguna dengan menggunakan *web browser*. Pada penelitian ini digunakan *random user agent* untuk menyamarkan proses *fetching* dari *web server* target, sehingga *web server* mengenali sebagai pengguna yang berbeda-beda.

Implementasi Wordpress Plugin

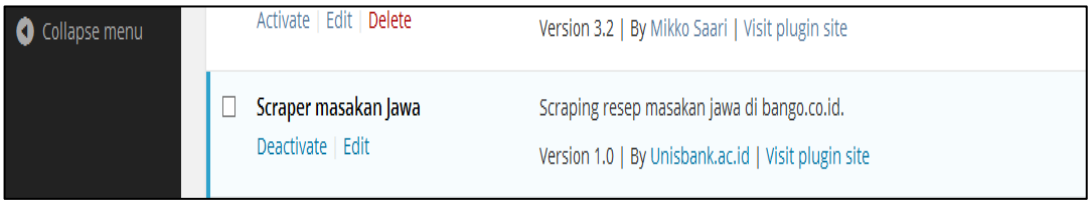
Setelah fungsi selesai ditulis maka untuk mempermudah penggunaan dan agar terintegrasi dengan wordpress maka struktur program fungsi *web scraping* diubah menjadi struktur *plugin* pada wordpress. Struktur program *plugin* pada wordpress mengharuskan ditambahkannya *header remark* yang berfungsi untuk memuat informasi seputar plugin tersebut.

Instalasi Wordpress

Proses instalasi wordpress dapat dilakukan dengan bantuan utilitas fantastico pada website dengan dukungan Cpanel. Apabila tidak terdapat ultilitas fantastico, kode sumber wordpress dapat diunduh pada URL http://wordpress.org/latest.zip. Kode sumber diekstrak pada direktori sesuai dengan kebutuhan. Instalasi dilanjutkan dengan menjalankan *script* http://[namadomain]/wp-admin/install.php. Setelah semua proses instalasi diselesaikan maka situs telah terinstal wordpress dan siap digunakan.

Instalasi Plugin Web Scraping

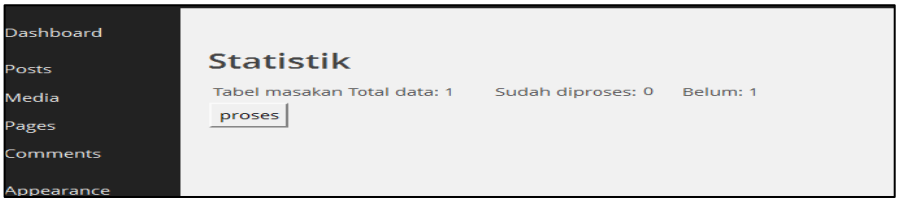
Plugin pada wordpress diinstal melalui menu administrator di URL http://[namadomain]/wp-admin/ setelah terlebih dahulu memasukan username dan password untuk administrator. Plugin dipasang pada menu *plugins-add new*. *Plugin* dapat diunggah ataupun langsung diunduh dari *repository* wordpress. Pada penelitian ini plugin tidak terdapat di *repository* wordpress, sehingga digunakan menu *upload* untuk menambahkan *plugin*. *Plugin* akan aktif setelah diaktifkan dengan mengklik *URL activate* dibawah nama *plugin*. Pada gambar 5 diperlihatkan tampilan layar daftar *plugin* yang terpasang di wordpress.



Gambar 5. Tampilan layar daftar plugin wordpress

Pengujian Memasukan Alamat URL Awal

Pengujian pertama dilakukan dengan URL pertama pada tabel masakan yaitu URL http://www.bango.co.id. URL lainnya akan secara otomatis diambil dari *link-link* pada halaman yang di-*scrap*. Pada gambar 6 diperlihatkan tampilan awal pada saat proses *scraping* dilakukan oleh admin. Pada gambar 7 diperlihatkan halaman hasil *scraping* pada alamat URL target.



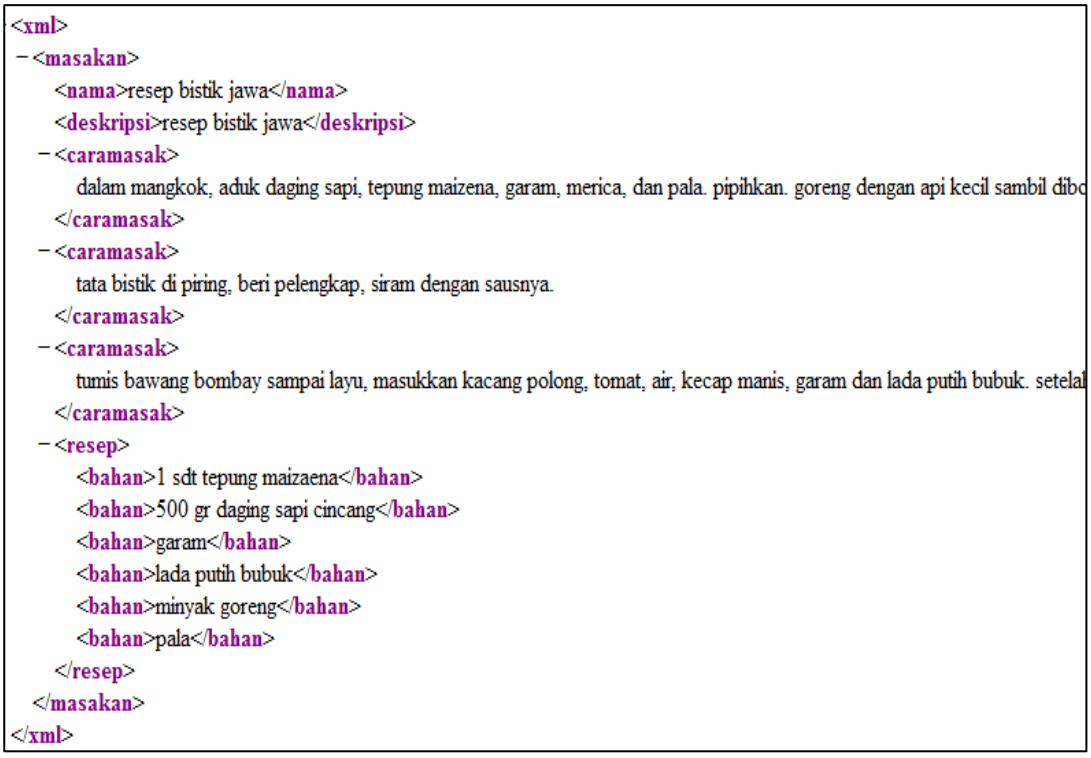
Gambar 6. Tampilan awal layar proses *scraping*

Gambar 7. Tampilan layar proses *scraping*

Pengujian Pengambilan File XML Resep Masakan

Proses pengambilan **XML** resep melalui URL utama http://masakbagus.com. Diikuti dengan parameter resep berisi nama resep. Contoh *request* adalah: http://masakbagus.com/?resep=resep bistik jawa. Maka akan dihasilkan resep masakan bistik jawa

dalam bentuk XML yang diperoleh dari pemrosesan URL tersebut, seperti yang terlihat pada gambar 8.



Gambar 8. Tampilan layar data XML

Kesimpulan

Berdasarkan hasil penelitian ini, maka dapat disimpulkan:

1. Teknik *web scraping* dengan metode regular ekspresi dapat digunakan untuk mengambil konten resep pada situs http://www.bango.co.id.
2. Penyimpanan resep ke dalam basisdata, mempermudah transformasi data ke bentuk lainnya.
3. Transformasi data dari mysql ke file berformat XML dapat dilakukan *on the fly* sehingga user seperti sedang mengakses file XML langsung.
4. Hasil dari *scraping* dapat dijadikan korpus untuk sistem temu kembali informasi.

Daftar pustaka

Bing, Liu, 2011, *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*, Second Edition, Springer.

Chaudhari, P. A. dan Paikrao, R. L., 2012, *Web Data Extraction,* International Journal of Computer Applications, 13-17.

http://w3c.org, diakses tanggal 12 Maret 2014.

Jennings, Frank dan Yates, John, 2008, *Scrapping over data: are the data scrapers' days numbered?*, Journal of Intellectual Property Law & Practice, 1-10.

Utomo, Mardi Siswo, 2013, *Web Scraping Pada Situs Wikipedia Menggunakan Metode Ekspresi Regular,* Jurnal Teknologi Informasi DINAMIK Volume 18, No.2, 153-160.