LONTAR KOMPUTER VOL. 8, NO. 3, DESEMBER 2017 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2017.v08.i03.p03 **e-ISSN 2541-5832**

Ekstrak Hirarki Data Dari Situs Web A-Z Animals Menggunakan Web Scraping

I Putu Arditya Darmawan[1], I Nyoman Piarsa[2], I Putu Arya Dharmaadi[3]

Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Udayana
Kampus Unud, Bukit Jimbaran, Bali, Indonesia
[1]putuarditya@gmail.com
[2]manpits@unud.ac.id
[3]aryadharmaadi@unud.ac.id

Abstrak

A-Z Animals merupakan sebuah website yang menyajikan data mengenai Kingdom Animalia. Data Kingdom Animalia memiliki hirarki atau tingkatan yang disebut dengan tingkat takson, yang dimulai dari kingdom hingga species. Permasalahan yang dihadapi adalah data yang terdapat pada website tersebut dapat digunakan kembali untuk kepentingan lain, seperti membuat kamus, media pembelajaran dan lain-lain, namun diperlukan waktu yang cukup lama untuk memasukkan data ke database karena data yang terlalu banyak dan kompleks. Solusi dari permasalahan tersebut adalah membuat aplikasi yang dapat secara otomatis mengambil data dari website untuk mempercepat pengumpulan data. Web Scraping merupakan metode untuk mengambil dokumen sebuah website dari internet, yang berupa HTML, selanjutnya dilakukan analisis untuk diambil data tertentu dari dokumen tersebut. Hasil pengujian yang telah dilakukan menunjukkan bahwa aplikasi dapat mengambil konten atau data yang diperlukan dari website a-z-animal.com. Aplikasi membutuhkan waktu rata-rata untuk memproses satu buah halaman a-z-animal.com adalah sekitar 16.13 detik.

Kata kunci: *Web Scraping, Kingdom Animalia, PHP, Ekstraksi Data.*

Abstract

A-Z Animals is a website that presents data about Kingdom Animalia. The Kingdom Animalia data has a hierarchy or level called the taxon level, which starts from kingdom to species. The problems encountered are the data contained on the website can be reuse for other purposes, such as creating dictionaries, learning media and others, but it takes a long time to enter data into the database due to the many and the complexity of the data. The solution of the problem is to create an application that can automatically retrieve data from the website to speed up data collection.Web Scraping is a method to retrieve documents from a website from the internet, in the form of HTML, next analyzed to retrieve certain data from the document. The results of tests sowed applications can retrieve content or data required from the website a-z-animal.com. The application takes an average time to process one page of a-z-animal.com is about 16.13 seconds.

Keywords: *Web Scraping, Kingdom Animalia, PHP, Data Extraction.*

1. Pendahuluan

Informasi memiliki kaitan yang erat dengan kehidupan masyarakat pada zaman sekarang. Teknologi yang berkembang sekarang mendorong informasi dapat diterima dengan mudah dan cepat. Teknologi yang berkembang dengan pesat sekarang adalah internet. Menurut pakar internet Onno W. Purbo, Internet merupakan sebuah media yang dapat digunakan sebagai sarana untuk saling bertukar informasi, baik berupa *web*, VoIP atau E-mail yang merupakan aplikasi dari internet [1]. Internet dapat mempermudah siapapun dalam pencarian informasi yang diinginkan.

Internet dapat digunakan dalam proses pengumpulan informasi, contohnya adalah *search engine* milik Google yang dapat membantu menjelajah di internet dengan mengumpulkan informasi dari

website. *Search engine* melakukan proses pengumpulan data dari berbagai *website* menggunakan *bot* secara periodik [2].

Data yang terdapat pada sebuah *website* dapat diolah dan digunakan kembali untuk kepentingan lain, seperti membuat kamus, media pembelajaran dan masih banyak lagi. Pembuatan media pembelajaran klasifikasi mahkluk hidup [3], yang memerlukan data spesies yang banyak sebagai materi dari media pembelajaran tersebut [4]. Data klasifikasi mahkluk hidup memiliki struktur yang bertingkat atau hirarki yang disebut dengan tingkat taksonomi. Tingkat takson dimulai dari *kingdom* hingga spesies. Data taksonomi yang diperlukan dapat diambil dari internet secara manual, tetapi akan membutuhkan waktu yang cukup lama untuk memproses data tersebut.

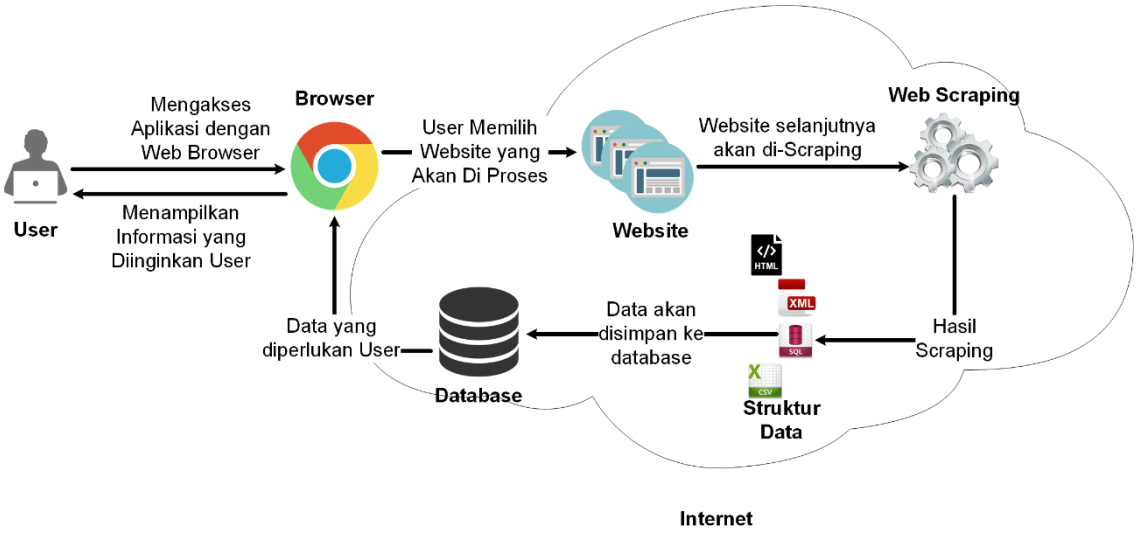
Data dari *website* dapat dikumpulkan dengan banyak cara selain diproses manual, contohnya dengan menggunakan Wget. GNU Wget atau Wget adalah sebuah paket *software* gratis yang berfungsi untuk mengambil *file* atau dokumen dengan menggunakan protokol HTTP, HTTPS dan FTP. Wget adalah sebuah *tool* yang berbasis *command line* atau menggunakan baris perintah untuk menjalankannya. Pengambilan data yang dilakukan Wget adalah mengunduh dokumen dari halaman *website* secara penuh, untuk pengambilan data yang lebih spesifik atau mengambil bagian tertentu saja dari sebuah *website* dapat menggunakan *web scraping* [5],[6].

Penelitian mengenai pengambilan data dari *website* telah banyak dilakukan, contohnya penelitian yang dilakukan oleh Utomo yaitu "*Web Scraping* pada Situs Wikipedia menggunakan Metode *Ekspresi Regular*" [7]. Aplikasi yang dibangun dan ditanam pada *web server* yang terkoneksi dengan jaringan internet. Aplikasi berjalan menggunakan *service* http dengan format transaksi data html, sehingga aplikasi dapat dibuka menggunakan terminal yang terkoneksi ke jaringan komputer dan mempunyai *browser web*. *User* dapat melihat dokumen yang telah diekstrak dalam bentuk artikel dalam wordpress. Komputer *Server* berfungsi sebagai *web server* yang telah terpasang Wordpress. *Web server* mengambil halaman *web* dari wikipedia.org kemudian mengekstrak konten utama dari halaman tersebut dan menyimpannya kedalam bentuk artikel di Wordpress. Penelitian lainnya dilakukan oleh Josi dengan judul "Penerapan Teknik *Web Scrapping* pada Mesin Pencari Artikel Ilmiah" [2]. Aplikasi yang dibuat berupa *web base* yang diimplementasikan dengan metode *web scraping* pada aplikasi yang telah dibuat, hasil dari pencarian disimpan ke dalam tabel menggunakan *database* MySQL.

Latar belakang tersebut yang mendorong melakukan penelitian ini. Penelitian ini berfokus pada proses pengambilan data pada *wesite* A-Z Animals dan memanfaatkan model data tree untuk menyimpan tingkat takson yang ada. Model tree akan mempermudah dalam pemodelan dari sistem klasifikasi yang ada.

2. Metodologi Penelitian

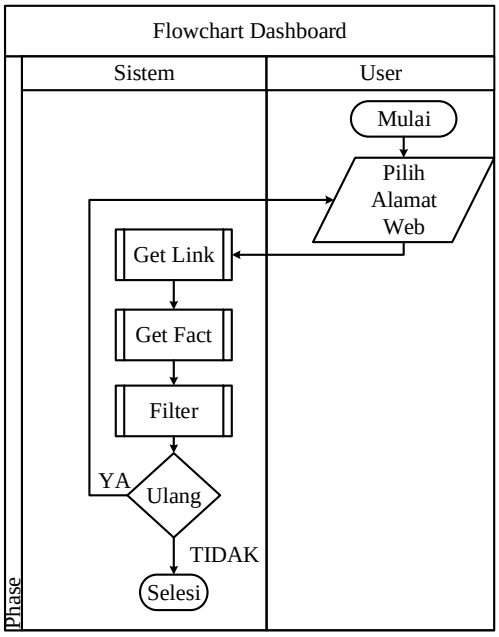
2.1. Gambaran Umum



Gambar 1. Gambaran Umum Sistem

Gambar 1 adalah gambaran umum dari aplikasi. Pertama *user* membuka *web browser* untuk mengakses aplikasinya, selanjutnya *user* memilih *web* yang diinginkan untuk diambil datanya. *Web* yang dipilih tersebut melanjutkan proses selanjutnya yaitu *scraping*, dalam proses *scraping* data yang diinginkan oleh *user* diekstrak dari *web* tersebut. Data yang berhasil diekstrak dari *web* tersebut disimpan di dalam *database*. Data yang disimpan tersebut digunakan oleh *user* untuk mendapatkan informasi yang diinginkan.

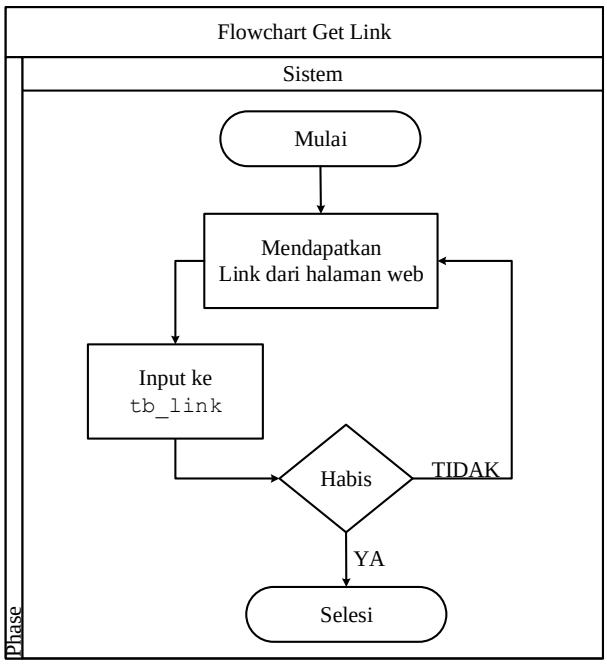
2.2. *Flowchart*



Gambar 2. *Flowchart Dashboard*

Gambar 2 flowchart ini menggambarkan alur kerja dari aplikasi. Terdapat tiga buah sub proses yaitu *Get Link*, *Get Fact* dan *Filter* yang dijelaskan sebagai berikut.

a. *Get Link*

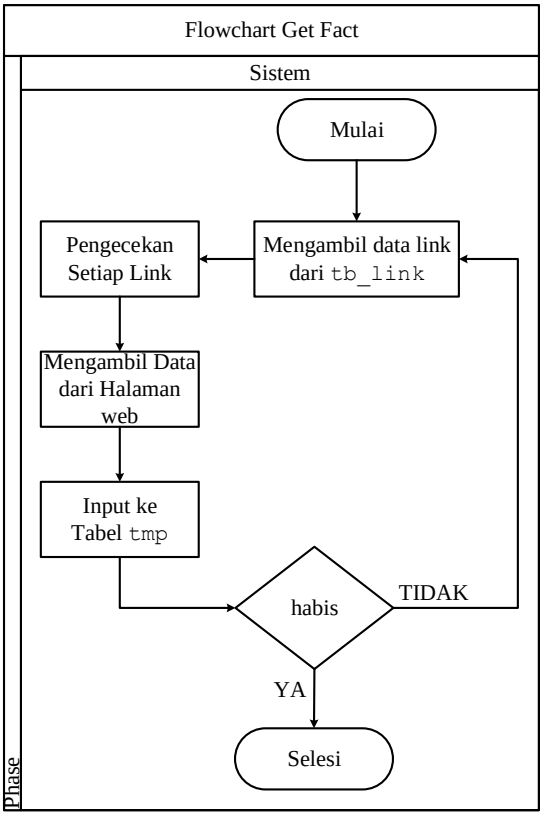


Gambar 3. *Flowchart Get Link*

Gambar 3 merupakan tampilan *flowchart* sub program *Get Link*. *Flowchart* ini menggambarkan alur kerja dari sub program *Get Link* yang berfungsi untuk mengidentifikasi halaman *web* yang dipilih *user* untuk mendapatkan *link* yang menghubungkan ke bagian info *animalia* pada *web* tersebut. *Link* yang berhasil didapat disimpan dalam database yang selanjutnya *link* tersebut digunakan pada sub program *Get Fact*.

b. *Get Fact*

Get Fact merupakan sub proses yang berfungsi untuk mengambil *fact* atau data yang diinginkan dari *website*. Alur proses dari *Get Fact* dapat dilihat pada Gambar 4.

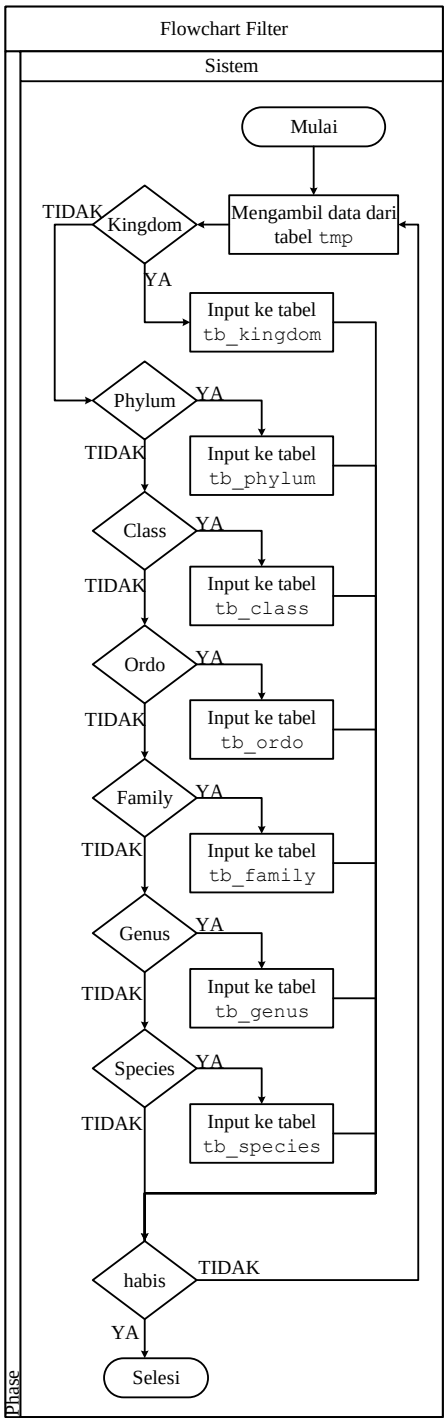


Gambar 4. *Flowchart Get Fact*

Gambar 4 merupakan tampilan *flowchart* sub program *Get Fact*. *Flowchart* ini menggambarkan alur kerja dari sub program *Get Fact* yang berfungsi untuk mengidentifikasi *link* yang telah disimpan sebelumnya pada *database*. *Link* tersebut selanjutnya diproses satu per satu dan diidentifikasi setiap halamannya, untuk mendapatkan data berupa data *animalia* yang ada pada halaman *web* tersebut. Data yang berhasil didapatkan selanjutnya disimpan dalam *database*.

c. *Filter*

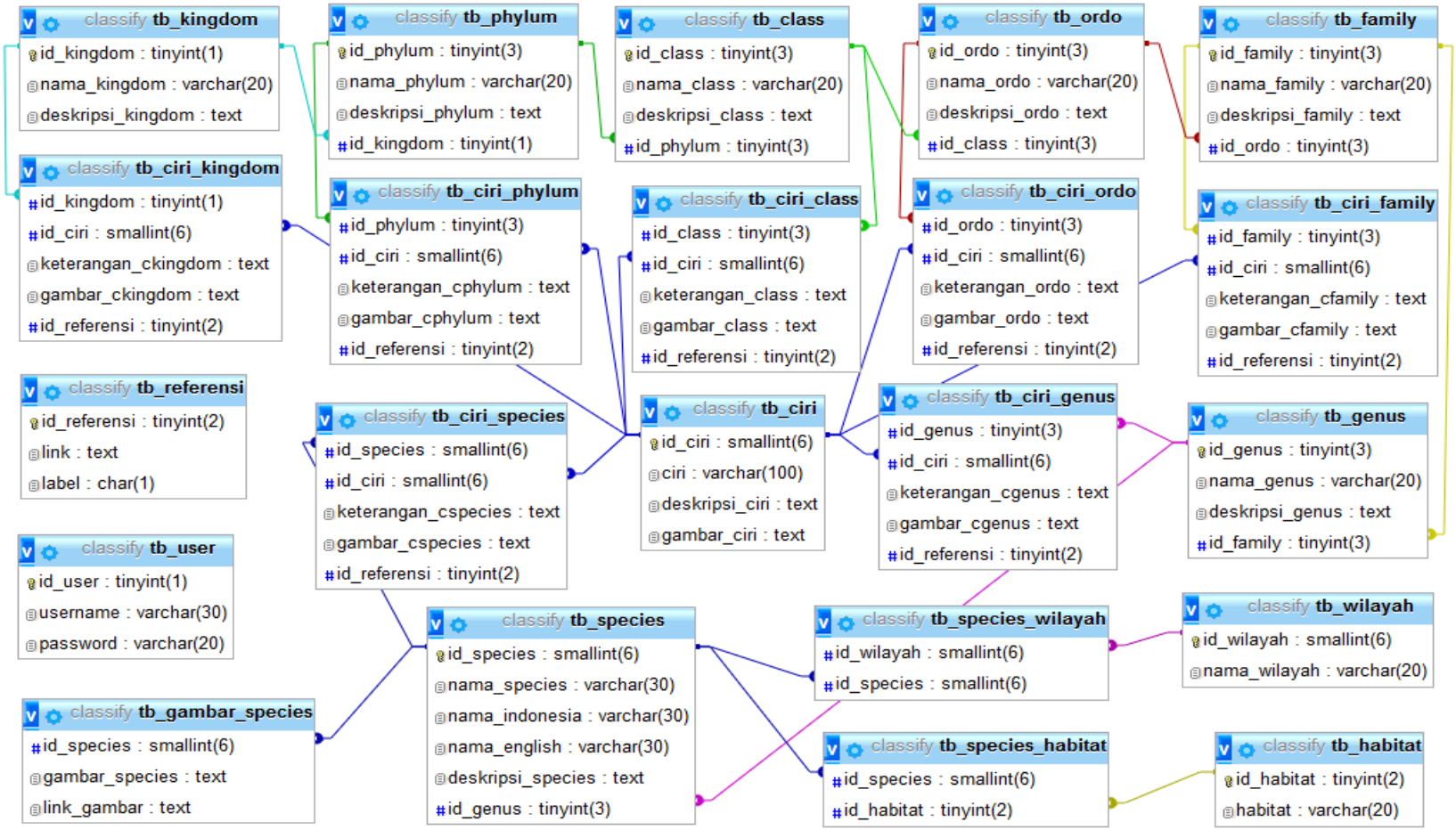
Filter merupakan sub proses yang berfungsi untuk melakukan *mapping* terhadap data yang berhasil diambil dari proses *Get Fact* sebelumnya. Data yang tersimpan pada proses *Get Fact* diperiksa satu per satu dalam proses *filter*. Data yang memenuhi syarat yang telah ditentukan dalam proses *filter* dimasukkan ke dalam *database*, sedangkan yang tidak memenuhi syarat maka dilewati. Alur kerja dari sub proses *filter* dapat dilihat pada Gambar 5.



Gambar 5. *Flowchart Filter*

2.3. *Relationship* tabel

Relationship tabel merupakan gambaran yang menunjukkan hubungan antara tabel-tabel yang telah dirancang sebelumnya. *Relationship* tabel dapat dilihat pada Gambar 6.



Gambar 6. *Relationship tabel*

3. Kajian Pustaka

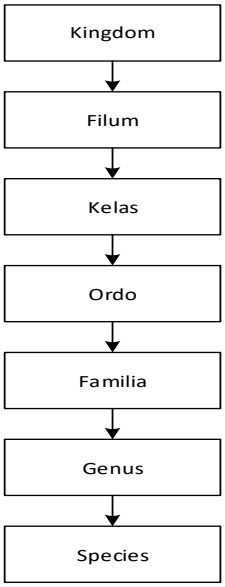
3.1. Klasifikasi

Sistem klasifikasi makhluk hidup terus berkembang hingga saat ini, karena adanya penemuan-penemuan baru yang dikembangkan oleh manusia. Sistem klasifikasi makhluk hidup bermula pada abad ke-19 sampai 20 masih menggunakan sistem dua *kingdom*, yaitu dunia tumbuhan (*Plantarum*) dan dunia hewan (*Animalia*). Penelitian yang dilakukan oleh Michael A. Ruggiero dan timnya memecah *kingdom* menjadi 7 bagian yang sebelumnya archae dan bacteria menjadi satu kini dipisah menjadi *kingdom* yang berbeda. Sistem klasifikasi 7 *kingdom* terdiri atas Kingdom Bacteria, Kingdom Archaea, Kingdom Protozoa, Kingdom Chromista, Kingdom Fungi, Kingdom Plantae dan Kingdom Animalia yang [8]. Klasifikasi adalah cara untuk melakukan pengelompokan terhadap makhluk hidup berdasarkan ciri-ciri tertentu. Tujuan dari klasifikasi adalah:

1. Melakukan pengelompokan pada makhluk hidup berdasarkan ciri-ciri yang dimiliki;
2. Menjelaskan mengenai ciri-ciri dari suatu jenis makhluk hidup agar dapat membedakan dengan jenis yang lainnya;
3. Mencari hubungan kekerabatan dari makhluk hidup yang ada;
4. Memberi nama kepada makhluk hidup yang tidak memiliki nama sebelumnya.

3.2. Tingkat Takson

Klasifikasi terdiri atas beberapa tingkatan, mulai dari kelompok besar, kemudian dibagi menjadi beberapa kelompok kecil, selanjutnya kelompok kecil dibagi menjadi beberapa kelompok kecil lagi sehingga terbentuk kelompok-kelompok yang lebih kecil yang hanya mempunyai anggota satu jenis makhluk hidup.



Gambar 7. Tingkatan Takson pada Kingdom Animalia

Gambar 7 merupakan tingkatan takson dari Kingdom Animalia. Takson tersebut tersusun dari tingkat tertingginya yaitu *kingdom* hingga yang terendah spesies, semakin tinggi tingkatan dari takson, maka persamaan ciri yang dimiliki akan semakin umum. Tingkatan takson yang semakin rendah, maka kesamaan ciri yang dimiliki makhluk hidup semakin khusus.

3.3. Struktur Data *Tree*

Metode *tree* atau pohon adalah sejumlah *node* yang berhubungan secara hirarkis dimana suatu *node* pada suatu hirarki merupakan cabang dari *node* dengan hirarki yang lebih tinggi dan juga memiliki cabang ke beberapa *node* lainnya dengan hirarki yang lebih rendah [9]. Metode *tree* dalam ilmu komputer adalah suatu struktur data yang digunakan secara luas yang menyerupai

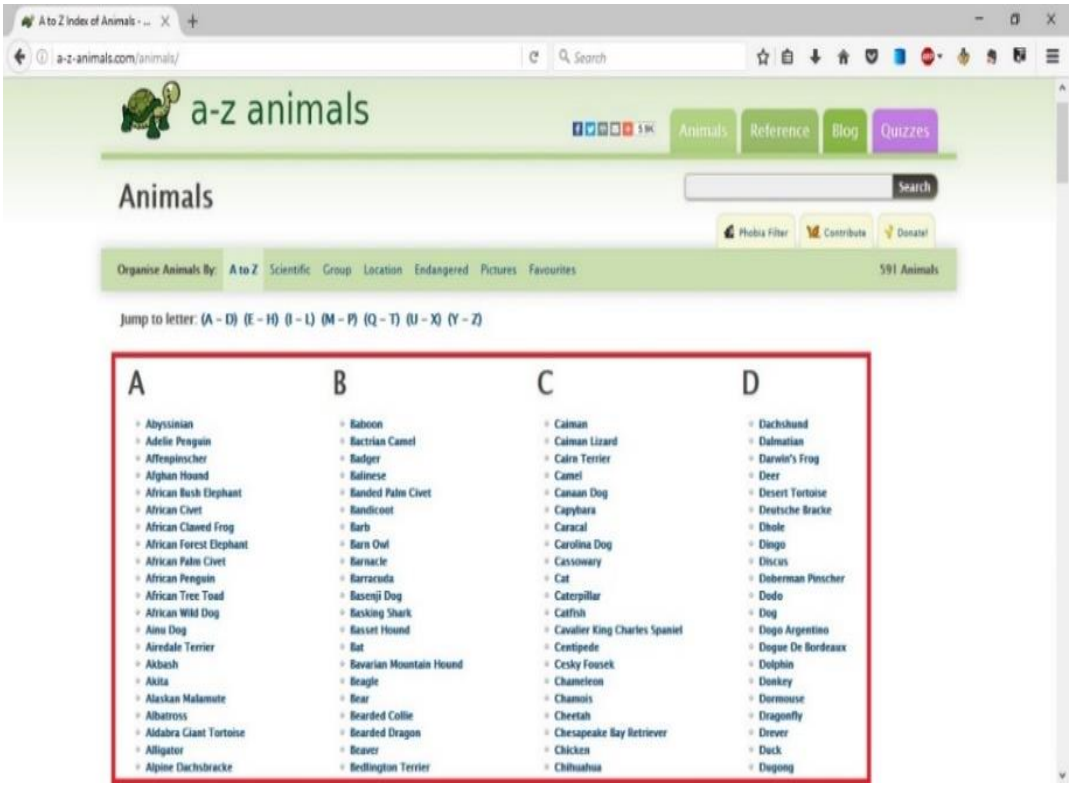
struktur pohon dengan sejumlah simpul yang terhubung [10]. Contoh apllikasi yang menggunakan metode *tree* adalah Sistem Informasi Upacara yadnya Berbasis Android [11].

3.4. Web Scraping

Web Scraping merupakan metode untuk mengambil dokumen sebuah *website* dari internet, yang berupa HTML maupun XHTML dan selanjutnya dilakukan analisis untuk diambil data tertentu dari dokumen tersebut. Data yang diambil dengan *web scraping* seperti *link*, gambar, maupun berita yang terdapat dalam sebuah *website* [2].

4. Hasil dan Pembahasan

Hasil dan pembahasan memaparkan mengenai hasil analisa dan pengujian pada aplikasi yang telah dikembangkan.



Gambar 8. Halaman a-z animals

Gambar 8 merupakan tampilan dari halaman *animals* dari *website* a-z-animals. Bagian yang diberi kotak merah adalah bagian yang dilakukan proses pengambilan data. Seluruh *link* pada bagian yang diberi kotak merah itu diambil dan disimpan ke dalam *database* oleh program.

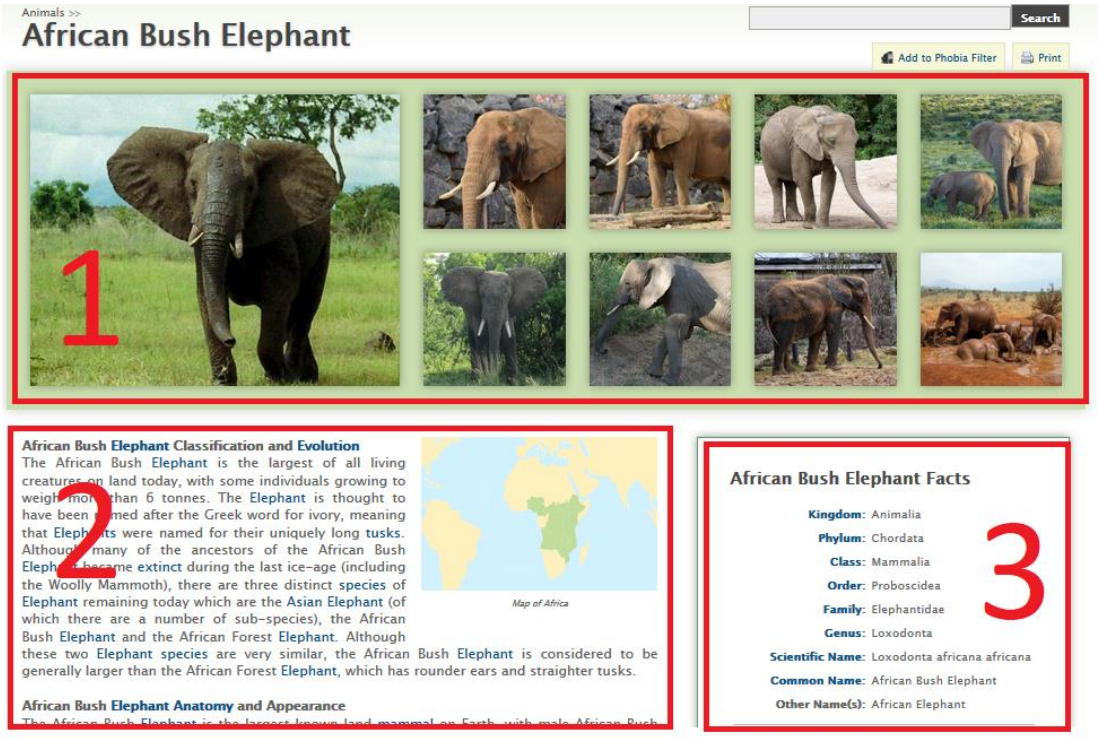
Show 10 entries Search:

NO	LINK	STATUS	ACTION
1	http://www.catalogueoflife.org/col/details/species/id/b481959b65f831cf3fe888a052c74ad7/source/tree	Checked	Edit | Delete
2	http://www.catalogueoflife.org/col/details/species/id/e5850208ceff1e76bb48ee50ab2394d5/source/tree	Checked	Edit | Delete
3	http://www.catalogueoflife.org/col/details/species/id/9b5442332562286347009e34977a39bf/source/tree	Checked	Edit | Delete
4	http://www.catalogueoflife.org/col/details/species/id/93a09f310f8273c81a556f4f78af896f/source/tree	Uncheked	Edit | Delete
5	http://www.catalogueoflife.org/col/details/species/id/222a428261b75966c0d5de31d95f5419/source/tree	Uncheked	Edit | Delete
6	http://www.catalogueoflife.org/col/details/species/id/1aa52e26fce7357b48d6d8a0b03c9ba1/source/tree	Uncheked	Edit | Delete
7	http://www.catalogueoflife.org/col/details/species/id/1c62a268fe49080730d30c447de258fd/source/tree	Uncheked	Edit | Delete
8	http://www.catalogueoflife.org/col/details/species/id/0fbaebc98ce5b28118ee36bd26096ced/source/tree	Uncheked	Edit | Delete
9	http://www.catalogueoflife.org/col/details/species/id/ca52f0aa78abb2a453590c004574a908/source/tree	Uncheked	Edit | Delete
10	http://www.catalogueoflife.org/col/details/species/id/ece7cacc050c03039727dcaf5f72b24b/source/tree	Uncheked	Edit | Delete

Showing 1 to 10 of 201 entries Previous 1 2 3 4 5 ... 21 Next

Gambar 9. Hasil Pengambilan *Link*

Gambar 9 menampilkan hasil dari *link* yang telah disimpan setelah proses pengambilan *link* selesai. Proses selanjutnya melakukan pengecekan terhadap *link* yang telah tersimpan di dalam *database*.



Gambar 10. Tabel Facts a-z animals

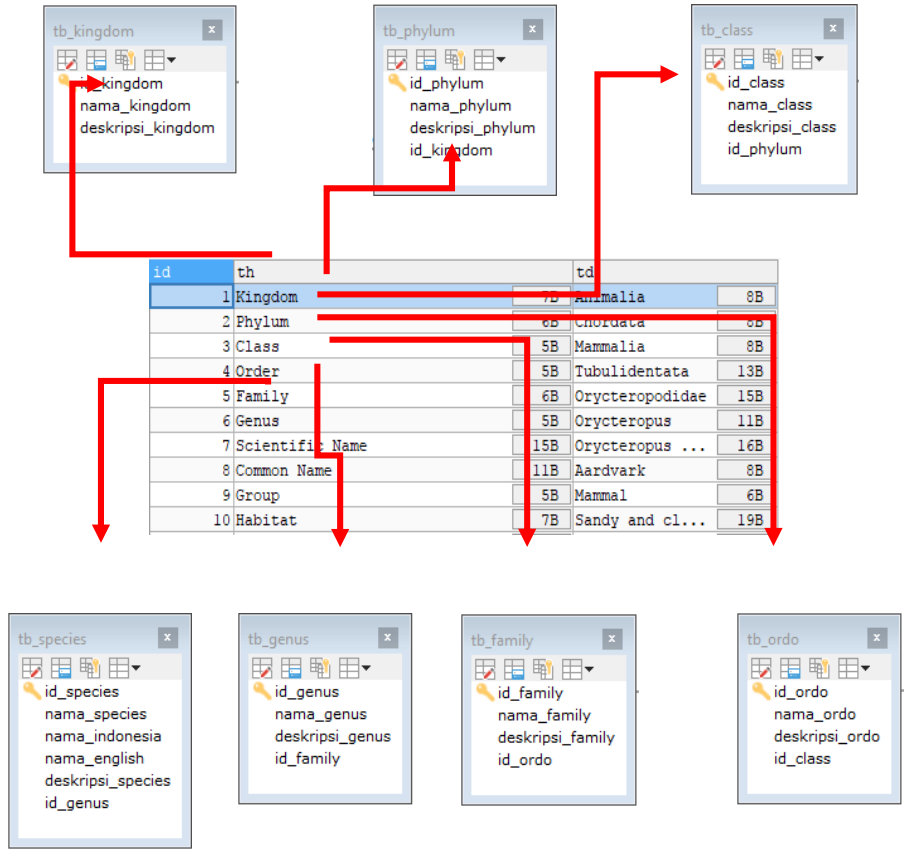
Gambar 10 merupakan tampilan dari halaman yang dilakukan proses *scraping*. Halaman ini diambil dari *link* yang tersimpan dalam *database* dari proses sebelumnya. Program membaca *link* tersebut satu per satu dan menampilkan halaman seperti Gambar 10. Terdapat tiga buah kotak merah pada halaman tersebut, kotak merah tersebut menunjukkan bagian yang diambil datanya.

Kotak nomor satu mengambil data berupa *link* dari gambar yang di tampilkan, kotak nomor dua mengambil deskripsi dari *spesies* tersebut dan kotak nomor tiga mengambil data berupa fakta dari *spesies* tersebut. Bagian tersebut yang diproses dan data yang berhasil diambil dimasukkan ke dalam *database*.

id	th		td	
1	Kingdom	7B	Animalia	8B
2	Phylum	6B	Chordata	8B
3	Class	5B	Mammalia	8B
4	Order	5B	Tubulidentata	13B
5	Family	6B	Orycteropodidae	15B
6	Genus	5B	Orycteropus	11B
7	Scientific Name	15B	Orycteropus ...	16B
8	Common Name	11B	Aardvark	8B
9	Group	5B	Mammal	6B
10	Habitat	7B	Sandy and cl...	19B

Gambar 11. Hasil *Scraping*

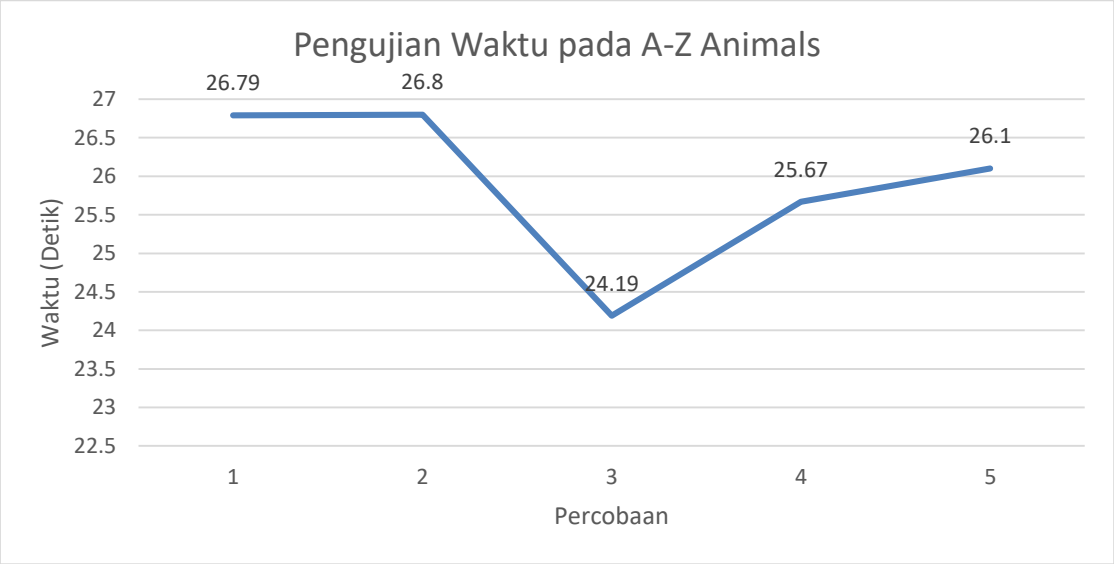
Gambar 11 merupakan hasil dari proses scraping yang telah dilakukan, dari Gambar 11 menampilkan data yang berasal dari halaman *web* seperti yang ditampilkan pada Gambar 10 pada kotak berwarna merah.



Gambar 12. Hasil *Scraping*

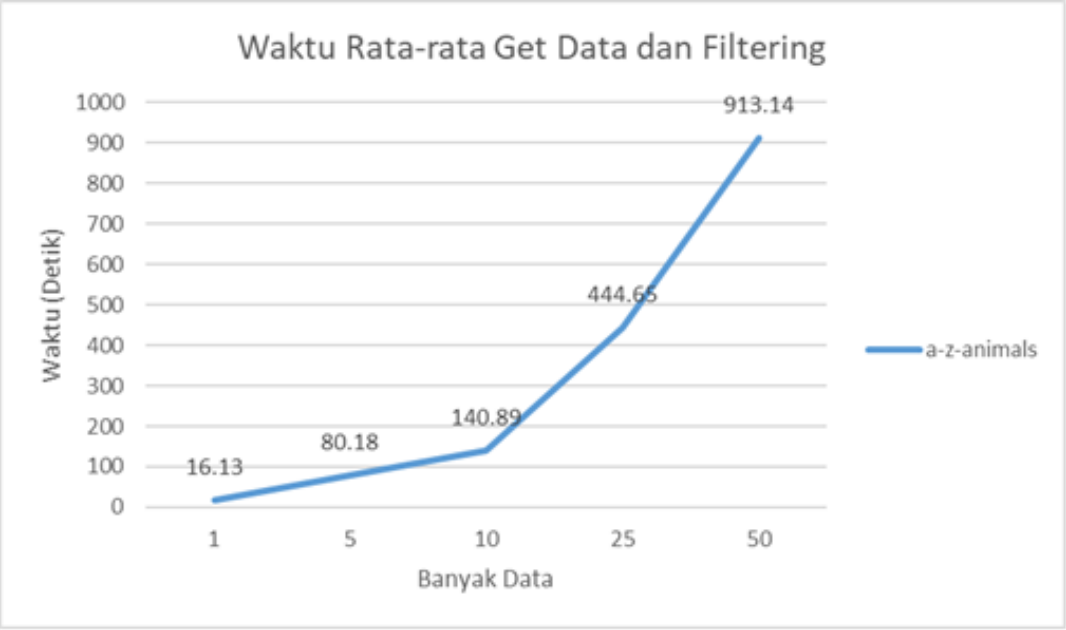
Gambar 12 menjelaskan mengenai *mapping* terhadap data yang telah berhasil disimpan ke dalam *database*. *Mapping* dilakukan jika proses pengambilan data sebelumnya sudah selesai. Data yang tersimpan dalam tabel tersebut dipindahkan ke masing-masing tabel seperti yang terlihat pada Gambar 12. Data *Kingdom* dimasukkan ke dalam `tb_kingdom`, data *Phylum*

dimasukkan ke dalam `tb_phylum`, data *Class* dimasukkan ke dalam `tb_class`, data *Order* dimasukkan ke dalam `tb_ordo`, data *Family* dimasukkan ke dalam `tb_family`, data *Genus* dimasukkan ke dalam `tb_genus` dan data *Scientific Name* dimasukkan ke dalam `tb_species`.



Gambar 13. Grafik Pengujian Waktu Pengambilan Link

Gambar 13 merupakan waktu rata-rata yang didapatkan dari pengujian sebanyak lima kali. a-z-animals.com membutuhkan waktu rata-rata 25.91 detik dan dapat mengambil data sebanyak 626 buah.



Gambar 14. Waktu Rata-rata *Get* Data dan *Filtering* a-z-animal

Pengambilan data dan *filtering* dari *website* a-z-animals.com setelah melakukan pengujian sebanyak tiga kali mendapatkan waktu rata-rata seperti yang terdapat pada Gambar 14.

5. Kesimpulan

Pengambilan konten atau data dari sebuah *website* melalui beberapa tahapan. Tahapan pertama adalah mempelajari struktur HTML dari *website*, yang bertujuan untuk menentukan bagian *website* yang ingin diambil datanya. Tahap kedua adalah memahami teknik navigasi pada

website, untuk selanjutnya ditirukan pada aplikasi *web scraper* agar dapat melakukan pencarian terhadap data yang diinginkan. Tahap ketiga adalah membuat otomatisasi program berdasarkan informasi yang didapatkan dari tahap satu dan dua. Tahap keempat yaitu melakukan penyimpanan data yang berhasil didapatkan ke *database*. Data yang didapatkan dari website a-z-animal.com berupa data takson, deskripsi dari hewan dan gambar dari hewan tersebut. Aplikasi akan mencari data tersebut berdasarkan syarat yang telah ditentukan pada saat tahap mempelajari struktur HTML *website*. Pengujian yang telah dilakukan menunjukkan bahwa waktu yang diperlukan aplikasi dalam proses pengambilan data *link* a-z-animal.com adalah sekitar 25.91 detik dan data yang didapatkan sebanyak 626 buah. Aplikasi membutuhkan waktu rata-rata memproses satu buah halaman a-z-animal.com adalah sekitar 16.13 detik.

Daftar Pustaka

[1] B. A. Nandari and Sukadi, "Pembuatan Website Portal Berita Desa Jetis Lor," *IJNS,* vol. 3, no. 3, pp. 43-47, 2014.

[2] A. Josi, L. A. Abdillah, and Suryayusra, "Penerapan Teknik Web Scrapping pada Mesin Pencari Artikel Ilmiah," *Jurnal Sistem Informasi,* vol. 5, no. 2, pp. 159-164, 2014.

[3] I. D. G. W. Dhiyatmika, I. K. D. Putra, and N. M. I. M. Mandenni, "Aplikasi Augmented Reality Magic Book Pengenalan Binatang untuk Siswa TK," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi,* vol. 6, no. 2, pp. 120-127, 2015.

[4] Wamiliana, D. Kurniasari, and J. S. Nugraha, "Pembuatan Media Pembelajaran Pengenalan Tata Surya dan Exoplanet Dengan Menggunakan Unity untuk Sekolah Menengah Pertama," *Jurnal Komputasi,* vol. 1, no. 1, pp. 47-57, 2013.

[5] F. Polidoro, R. Giannini, R. L. Conte, S. Mosca, and F. Rossetti, "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation," *Statistical Journal of the IAOS,* pp. 165–176, 2015.

[6] M. A. Pise and P. J. Adhikari, "A Review: Data Extraction from multiple web databases," *IJRITCC,* vol. 3, no. 10, pp. 5930-5932, 2015.

[7] M. S. Utomo, "Web Scraping pada Situs Wikipedia menggunakan Metode Ekspresi Regular," *Jurnal Teknologi Informasi DINAMIK* vol. 18, no. 2, pp. 153-160, 2013.

[8] M. A. Ruggiero, D. P. Gordon, T. M. Orrell, N. Bailly, T. Bourgoin, R. C. Brusca*, et al.*, "A Higher Level Classification of All Living Organisms," *PLOS ONE,* pp. 1-54, 2015.

[9] I. G. B. A. Pinatih, A. A. K. Oka Sudana, and I. K. Adi Purnawan, "E-Banjar Bali, Population Census Management Information System of Banjar in Bali by Using Family Tree Method and Balinese Culture Law," *Journal of Theoretical and Applied Information Technology,* vol. 59, no. 2, pp. 411-420, 2014.

[10] A. A. K. Oka Sudana, I. W. G. M. Kepakisan, and N. K. D. Rusjayanthi, "Implementation of Tree Structure and Recursive Algorithm for Balinese Traditional Snack Recipe on Android Based Application " *International Journal of Interactive Mobile Technologies,* vol. 10, no. 4, pp. 43-47, 2016.

[11] I. M. W. Saputra, A. A. K. Oka Sudana, and I. M. Sukarsa, "Implementasi Struktur Data tree pada Sistem Informasi Upacara yadnya Berbasis Android," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi,* vol. 2, no. 1, pp. 326-334, 2014.