MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 1 | No. 2 | Halaman 76 - 84
DOI : https://doi.org/10.26760/mindjournal Juli 2016

Implementasi *Ontology* Pada *Web Crawler*

JASMAN PARDEDE, UUNG UNGKAWA, MUHAMMAD AKBAR BERNOVALDY

Jurusan Teknik Informatika, Fakultas Teknologi Industri
Institut Teknologi Nasional

E-mail : Jasmanpardede78@gmail.com

ABSTRAK

Web crawler adalah suatu program atau script otomatis yang bekerja dengan memprioritaskan ketentuan khusus untuk melakukan penjelajahan dan melakukan pengambilan informasi dalam halaman web yang ada di internet. Proses pengindeksan merupakan proses crawler yang memudahkan setiap orang dalam pencarian informasi Pada proses indexing tersebut dibangun dengan menggunakan metode ontology. Metode ontology merupakan sebuah teori tentang makna dari suatu objek dengan hubungan objek tersebut. Pada penelitian ini, metode ontology diterapkan dalam proses pengambilan data dan pengelompokkan data. Metode ontology memiliki proses, yaitu melakukan splitting terhadap objek dengan ketentuan relasi untuk mendapatkan sebuah objek ontology. Selanjutnya dilakukan crawling terhadap objek ontology tersebut untuk mendapatkan hasil crawling dengan ontology. Pengelompokkan data diproses berdasarkan objek yang telah didapat berdasarkan relasi ontology. Dari hasil penelitian dapat diambil kesimpulan, yaitu presentase objek relasi sesuai dengan relasinya adalah 100% dan kecepatan web crawler dengan ontology lebih cepat 56,67% dibanding dengan web crawler biasa.

Kata kunci : *web crawler, ontology, web archiving.*

ABSTRACT

Web crawler is a program or automated script which works by prioritizing specific provisions to browsing and retrieval of information in a web page on the internet. Ontology process are divided by three, is scraping, parsing and indexing. Process of indexing is crawling process that makes it easy everyone search of information. In the process of indexing is built using methods of ontology. Methods of ontology is a theory about the meaning of an object with the object relationship. In this study, the method ontology applied in the data collection process and grouping data. Ontology method has a process of conducting splitting of the object with the provisions of relations to get an object ontology. Furthermore, the crawling towards the ontology objects to get the result crawling with ontology. Grouping the data is processed by an object that has been obtained based on ontology relationships. From the research results can be concluded, that the percentage of object relations in accordance with the relation is 100% and the speed of a web crawler with ontology 56.67% faster than regular web crawler.

Keyword : *web crawler, ontology, web archiving*

1. PENDAHULUAN

Web crawler merupakan suatu program atau script otomatis yang relatif simple karena bekerja dengan metode tertentu untuk melakukan scan atau "crawl" ke semua halaman web di internet untuk membuat indeks dari data yang dicari[6]. *Web crawler* memiliki tiga proses, yaitu pembacaan halaman (*scraping),* pemisahan kata *(parsing)* dan pengindeksan (*indexing).* Proses pengindeksan (indexing) merupakan proses *crawler* yang memudahkan setiap orang dalam pencarian informasi. Metode yang dapat digunakan untuk proses pengindeksan adalah metode *ontology.* Metode *ontology* merupakan sebuah metode tentang keterkaitan suatu objek dengan objek yang lain berdasarkan sebuah relasi tertentu.

Pada penerapan metode *ontology, crawler* hanya mengambil halaman yang harus diambil berdasarkan pengelompokkan kategori. Sehingga didapatkan sebuah batasan seperti pengambilan alamat web untuk *crawler* hanya dari web kompas dan detik, pengindeksan dilakukan pada *page* 1 dari *domain* dan *sub-domain,* data yang digunakan untuk kebutuhan adalah data pada Bulan Oktober 2016, *file* disimpan dalam bentuk .docx, pengambilan alamat web berdasarkan *element* yang sudah ditentukan dan studi kasus yang diambil merupakan pengarsipan dari sekali proses *crawling.*

Pada penelitian ini dibahas bagaimana data diindeks dengan metode *ontology,* kemudian dilakukan pencocokkan objek *ontology* dengan relasi *ontology* dan pengujian *respond time* pada *web crawler*. Sehingga didapat hasil implementasi *ontology* pada *web crawler* dengan parameter kecocokkan objek *ontology* dan *respond time* dari *web crawler* dengan *ontology*

2. METODOLOGI PENELITIAN

Web crawler menjelajahi halaman web berdasarkan alamat web yang telah diberikan, yaitu kompas.com atau detik.com. Hasil akhir dari *web crawler* sendiri diaplikasikan sebagai sistem pengarsipan berupa alamat web, judul web dan konten dari web tersebut. *Web crawler* memiliki tiga proses kerja, yaitu pembacaan halaman (*scraping),* pemisahan kata (*parsing)* dan pengindeksan (*indexing).*

Scraping merupakan proses awal dalam *web crawler* yang fungsinya untuk mengambil halaman web. *Scraping* berkaitan erat dengan pengindeksan, seperti untuk bagaimana mengembangkan teknik *scrape* yaitu dengan terlebih dahulu mempelajari dokumen HTML dari website yang dijelajahi dengan informasi yang diambil berupa tag HTML, tujuannya adalah informasi yang dikumpulkan setelah pembuat program belajar teknik navigasi akan diterapkan ke dalam aplikasi.[1] Dalam langkah pertamanya, *scraping* bekerja dengan mendapatkan *html* berdasarkan alamat web yang di-*input* oleh *user.* Dokumen *html* tersebut diuraikan dalam beberapa bagian dengan penandaan kata diantara tanda "<" dan ">". Tahap terakhir, pembacaan kata tersebut dibagi dalam *tag html, body* dan *head*, dimana sekumpulan kode dibaca dan dibagi dalam ketiga bagian tersebut.

Web crawler memisahkan kata dari keseluruhan halaman berdasarkan *link* yang diambil. Tiap *link* dalam halaman didefinisikan dengan sebuah penanda untuk pembacaan sebuah *link,* yaitu *element* "a href". Pemisahan kata sendiri dalam *web crawler* digunakan untuk pengambilan *link* atau pranala *link* tertentu.

Proses pengambilan data pada kata yang telah dipisah berdasarkan *element* "a href":

<a href= "http://megapolitan.kompas.com">

<a href= "http://nasional.kompas.com">
Hasil proses *parsing* ini adalah sebuah *link* (alamat web) yang ada pada dokumen html yang telah di-*scraping*

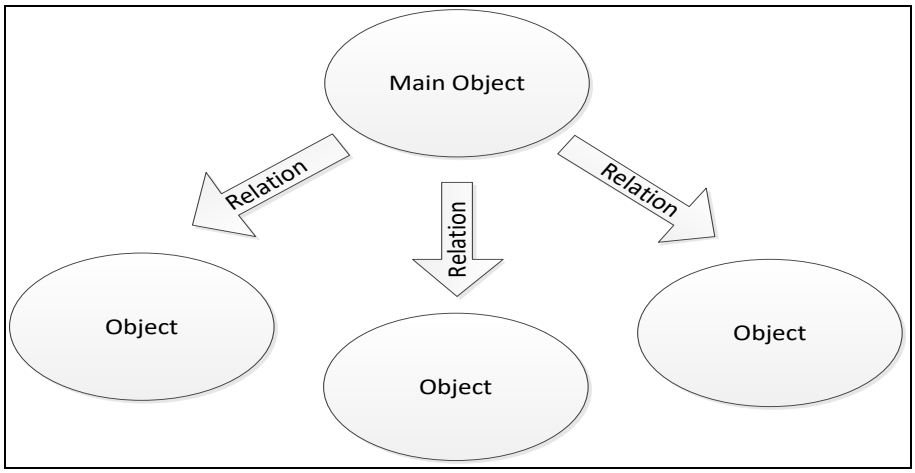
http://megapolitan.kompas.com
http://nasional.kompas.com

Indexing merupakan sebuah proses dimana metode *ontology* diterapkan. Pada penelitian ini, proses pengelompokkan *ontology* dilakukan berdasarkan pengelompokkan kategori. Kategori merupakan sebuah kelompok yang terdiri dari alamat web tertentu yang memiliki relasi sebagai bagian dari alamat web utama. *Ontology* sendiri diterapkan pada proses *indexing* karena proses *indexing* merupakan sebuah proses yang berfungsi untuk memudahkan penyimpanan informasi dalam proses pencarian di *search engine*. Pada proses *indexing* juga memudahkan penerapan metode *ontology* karena dalam proses ini pengelompokkan dilakukan dengan membedakan setiap bagian. Sehingga penerapan *ontology* yang membedakan sebuah makna berdasarkan ketentuan relasi dapat diterapkan.

Indexing dilakukan dengan *splitting* sebuah kata berdasarkan pemisahan tanda "/". Tahap kedua, pengecekkan kata yang telah dipisah dengan ketentuan relasi *ontology.* Setiap kata dari alamat web yang dipisah berdasarkan tanda "/" dicocokkan dengan kode kategori dan kode alamat web dari ketentuan relasi *ontology.* Kata-kata yang telah sesuai dikelompokkan berdasarkan objek *ontology* yang didapat dari ketentuan relasi. Sekumpulan kata tersebut yang berupa alamat web yang terindeks dengan pengelompokkan kategori adalah *indexing ontology.*

Ontology merupakan suatu teori atau metode tentang makna dari suatu objek, properti dari suatu objek, serta relasi objek tersebut yang mungkin terjadi pada suatu domain pengetahuan. *Ontology* dapat diartikan juga sebagai penjelasan sebuah konsep yang memiliki hubungan atau kaitan dari ilmu tertentu. Sebuah *ontology* terdiri dari sebuah daftar istilah terbatas dan hubungan diantara istilah-istilah[3]. *Ontology* biasanya disusun sebagai serangkaian konsep yang memiliki keterkaitan semantic. Semantik adalah cabang linguistik yang mempelajari secara khusus tentang arti, perubahan arti, dan prinsip hubungan antaran kata dan artinya. Sebuah *ontology* adalah spesifikasi yang eksplisit dan formal dari sebuah konseptualisasi.[7]

Adapun struktur *ontology* sederhana ditampilkan pada Gambar 1. Pada Gambar 1, *Main Object* merupakan sebuah istilah untuk objek yang menjadi sebuah pusat dari hubungan *ontology* atau istilah-istilah yang saling memiliki relasi. Relasi adalah sebuah keterkaitan atau hubungan yang menghubungkan dua objek atau lebih. Sedangkan *object* adalah istilah-istilah yang berelasi dari objek utama tersebut.



Gambar 1. Struktur Ontology

Proses penerapan *ontology* dalam *indexing* memerlukan beberapa kebutuhan data. Analisis kebutuhan data untuk metode *ontology* diproses menjadi sebuah kata kunci khusus dalam membantu pemrosesan *web crawler* dengan *ontology.* Kebutuhan data diambil berdasarkan pengumpulan data sampel yang diperoleh dari website kompas dan detik pada bulan Oktober 2016, dengan didapatkan kesimpulan, yaitu :

1. Data relasi *ontology* ditampilkan pada Tabel 1 merupakan sebuah kata kunci yang bekerja sebagai relasi dalam hubungan *ontology.* Kata kunci atau relasi tersebut didapat berdasarkan sebuah keterkaitan yang dimiliki objek utama. Keterkaitan yang dijadikan acuan pada penelitian ini adalah sebuah relasi berdasarkan "kategori" yang dimiliki oleh *website* yang menjadi objek utama.
2. Data kode judul, berupa inisialisasi kode yang digunakan untuk penentuan judul yang diambil dari kata atau bagian terakhir pada alamat web yang setiap kata atau bagiannya dipisah oleh "/".
3. Data *element* konten artikel yang ditampilkan pada Tabel 2, merupakan sebuah *element* yang digunakan untuk proses pengambilan artikel. Penentuan *element* ini ditentukan berdasarkan sebuah *element* yang menjadi wadah untuk konten artikel yang dimiliki oleh website yang menjadi objek utama.

Tabel 1. Data Relasi

Alamat Web	Kode Relasi
Kompas.com	xxxx.kompas.com
Detik.com	m.detik.com/xxxx

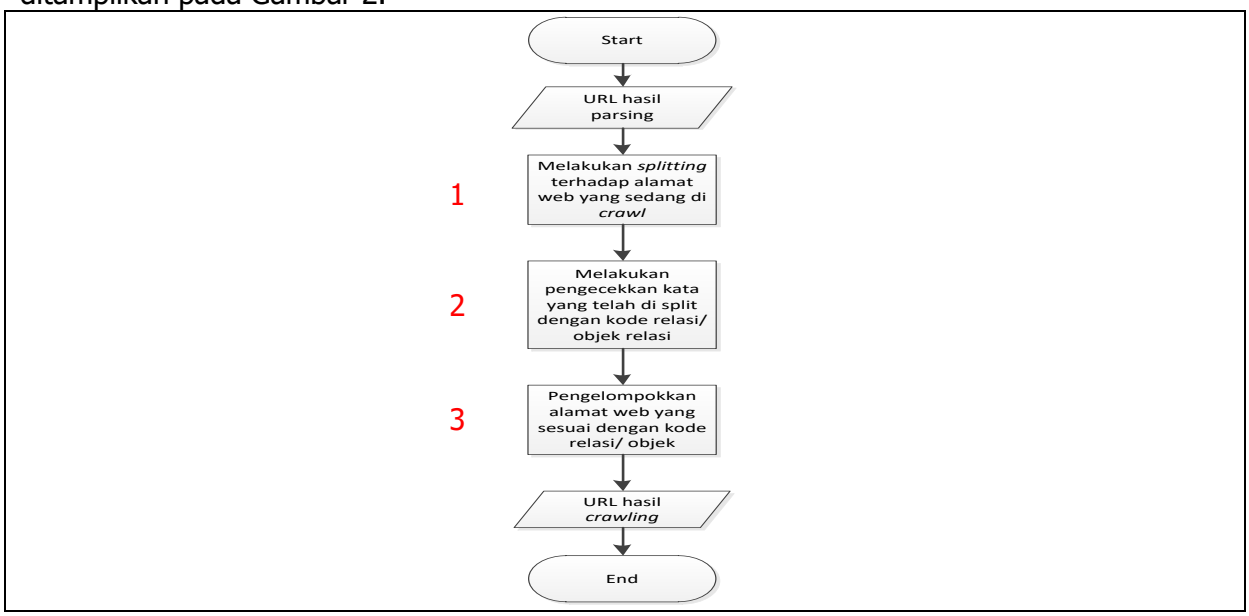
Tabel 2. Data *Element* Konten Artikel

No	Alamat Web	*Element* Konten Artikel
1.	KOMPAS.COM	div.kcm-read-text
2.	Detik.com	div.text_detail

3. ANALISIS DAN PEMBAHASAN

Proses *indexing* merupakan proses pengelompokkan data berdasarkan ketentuan relasi dan objek *ontology.* Pada proses *indexing* diterapkan metode *ontology* dengan ketentuan relasi

berdasarkan kategori dari alamat web utama (*main object).* Tahap pengindeksan *ontology* ditampilkan pada Gambar 2.



Gambar 2. Flowchart Proses *Indexing*

Pada proses *indexing* dilakukan 3 tahapan, yaitu :

1. Melakukan *splitting* pada alamat web yang sedang di *crawl. Splitting* yang dilakukan berdasarkan pemisahan tanda "/" dengan ketentuan relasi atau objek *ontology.*

Hasil *parsing* berupa sekumpulan alamat web pada objek "megapolitan" :

1. http://megapolitan.kompas.com/URL1
2. http://nasional.kompas.com/URL1
3. http://megapolitan.kompas.com/URL2
4. http://gramedia.com
5. http://megapolitan.kompas.com/URL3
6. http://nasional.kompas.com/URL2

Tahap *splitting* berdasarkan pemisahan "/" pada alamat web nomer 1 :

http:**//**megapolitan.kompas.com**/**URL1

Hasil *splitting* alamat web nomer 1 :

megapolitan.kompas.com

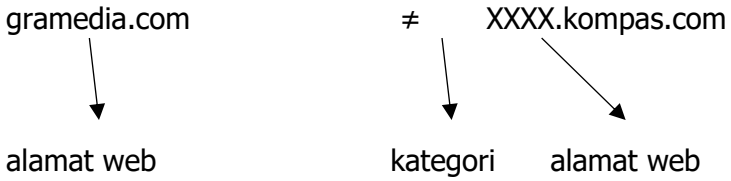
2. Pengecekkan kata yang telah di *split* dengan relasi *ontology* atau objek *ontology.* Pada tahap ini, dilakukan pengecekkan kata dari tahap *splitting* dengan ketentuan *ontology.* Jika hasil *crawl* sesuai dengan relasi atau objek *ontology*-nya, maka hasil *crawl* diambil dan disimpan. Jika tidak maka hasil *crawl* tidak diambil.

Tahap pengecekkan alamat web nomer 1 :

megapolitan.kompas.com = XXXX.kompas.com

kategori alamat web kategori alamat web

Tahap pengecekkan alamat web nomer 3 :



Hasil pengecekkan :

1. http://megapolitan.kompas.com/URL1
2. http://nasional.kompas.com/URL1
3. http://megapolitan.kompas.com/URL2
4. http://megapolitan.kompas.com/URL3
5. http://nasional.kompas.com/URL2

3. Tahap terakhir, dilakukan pengelompokkan hasil *crawl* yang telah diambil berdasarkan dengan relasi *ontology* atau objek *ontology.*

Tahap pengelompokkan alamat web pada objek "megapolitan" :

1. http://megapolitan.kompas.com/URL1
   ~~http://nasional.kompas.com/URL1~~
2. http://megapolitan.kompas.com/URL2
3. http://megapolitan.kompas.com/URL3
   ~~http://nasional.kompas.com/URL2~~

Hasil *indexing* pada alamat web dengan objek "megapolitan" :

1. http://megapolitan.kompas.com/URL1
2. http://megapolitan.kompas.com/URL2
3. http://megapolitan.kompas.com/URL3

Hasil proses *indexing* dengan *ontology* adalah alamat web yang telah dikelompokkan berdasarkan ketentuan relasi *ontology* dan pengelompokkan berdasarkan objek *ontology.* Berbeda dengan proses *indexing* tanpa *ontology* yang hanya melakukan pengelompokkan berdasarkan urutan pengambilan alamat web, dimana hasil yang disimpan tidak jauh berbeda dengan hasil *parsing*. Sehingga hasil dengan *ontology* sudah terstruktur berdasarkan objeknya.

3.1 Pengujian Kecocokkan Objek *Ontology*

Pada pengujian kecocokkan objek *ontology* ini bertujuan untuk melihat kesesuaian objek pada pengambilan objek berdasarkan relasi *ontology.* Objek-objek tersebut diuji berdasarkan kecocokan bentuk kode dari url tersebut, dimana objek atau url memiliki bentuk yang sesuai dengan kode objek relasi dan kode alamat web pada relasi *ontology*.

Setelah keadaan atau kondisi sebuah objek atau url sesuai dengan relasi yang telah ditentukan, maka objek atau url tersebut merupakan bagian dari objek *ontology.* Sekumpulan objek yang memiliki keterkaitan dengan objek utamanya. Objek-objek berelasi ini yang kemudian diproses dan di-*crawl* dalam tahap akhir *crawling* dengan *ontology.* Berikut analisis struktur *ontology* yang diuji untuk web kompas dan detik pada 21 Desember 2016, ditampilkan pada Tabel 3 dan Tabel 4.

Tabel 3. Pengujian Struktur *Ontology* Kompas.com

No	Url Relasi	Objek Relasi (xxxx)	Alamat Web (kompas)	Validitas
1.	http://nasional.kompas.com	Nasional	Kompas	Valid
2.	http://regional.kompas.com	Regional	Kompas	Valid
3.	http://megapolitan.kompas.com	Megapolitan	Kompas	Valid
4.	http://internasional.kompas.com	Internasional	Kompas	Valid
5.	http://olahraga.kompas.com	Olahraga	Kompas	Valid
6.	http://sains.kompas.com	Sains	Kompas	Valid
7.	http://bisniskeuangan.kompas.com	Bisniskeuangan	Kompas	Valid
8.	http://bola.kompas.com	Bola	Kompas	Valid
9.	http://tekno.kompas.com	Tekno	Kompas	Valid
10.	http://vik.kompas.com/	Vik	Kompas	Valid
11.	http://entertainment.kompas.com	Entertainment	Kompas	Valid
12.	http://otomotif.kompas.com	Otomotif	Kompas	Valid
13.	http://health.kompas.com	Health	Kompas	Valid
14.	http://female.kompas.com	Female	Kompas	Valid
15.	http://properti.kompas.com	Properti	Kompas	Valid
16.	http://travel.kompas.com	Travel	Kompas	Valid
18.	http://edukasi.kompas.com	Edukasi	Kompas	Valid
19.	http://kolom.kompas.com	Kolom	Kompas	Valid
20.	http://foto.kompas.com	Foto	Kompas	Valid
21.	http://video.kompas.com	Video	Kompas	Valid
22.	http://tv.kompas.com/	Tv	Kompas	Valid
23.	http://indeks.kompas.com/	Indeks	Kompas	Valid
24.	http://news.kompas.com/	News	Kompas	Valid

Tabel 4. Pengujian Struktur *Ontology* Detik

No	Url Relasi	Alamat Web (detik)	Objek Relasi (xxxx)	Validitas
1.	http://m.detik.com/news	Detik	News	Valid
2.	http://m.detik.com/finance	Detik	Finance	Valid
3.	http://m.detik.com/hot	Detik	Hot	Valid
4.	http://m.detik.com/sport	Detik	Sport	Valid
5.	http://m.detik.com/sepakbola	Detik	Sepakbola	Valid
6.	http://m.detik.com/inet	Detik	Inet	Valid
7.	http://m.detik.com/oto	Detik	Oto	Valid
8.	http://m.detik.com/wolipop	Detik	Wolipop	Valid
9.	http://m.detik.com/health	Detik	Health	Valid
10.	http://m.detik.com/travel	Detik	Travel	Valid
11.	http://m.detik.com/food	Detik	Food	Valid
12.	http://m.detik.com/tv	Detik	Tv	Valid
13.	http://m.detik.com/foto	Detik	Foto	Valid
14.	http://m.detik.com/pasangmata	Detik	Pasangmata	Valid
15.	http://detik.com/pilkadadki	Detik	Pilkadadki	Valid

Setelah pengujian telah dilakukan, maka berdasarkan pencocokkan setiap objek tersebut didapatkan hasil 100% objek *ontology* sesuai dengan relasi *ontology*.

3.2 Pengujian Respond Time Berdasarkan Kecepatan

Pengujian *respond time crawling* dengan *Ontology* merupakan pengujian dari *respond time* sistem utama yang telah dirancang. Pengujian *respond time crawling* dengan *Ontology* tersebut dilakukanmenggunakan parameter kecepatan Dalam melakukan kedua pengujian

tersebut dilakukan *crawling* dengan metode *Ontology* dan *crawling* tanpa menggunakan metode *Ontology* untuk medapatkan perbedaan hasil *crawling* menggunakan metode *ontology* dan tidak menggunakan metode *ontology.* Perbandingan *respond time* kecepatan *crawling* tersebut ditampilkan pada Tabel 5.

Tabel 5 Pengujian Respond Time *Ontology*

Pengujian	Crawl Ontology Kompas (s)	Crawl Kompas (s)	Crawl Ontology Detik (s)	Crawl Detik (s)
1	5	7	3	4
2	5	7	4	4
3	7	9	7	11
4	8	11	9	9
5	4	4	4	4
6	4	4	4	4
7	9	7	4	7
8	5	4	8	7
9	4	5	12	7
10	5	7	6	7
11	7	9	2	7
12	5	5	9	7
13	5	8	5	7
14	9	6	5	7
15	4	5	6	7

Lebih cepat Lebih lambat Kecepatan sama

Berdasarkan data pengujian yang telah dilakukan pada Tabel 5, maka didapatkan hasil 56,67% *web crawler* dengan *ontology* lebih cepat dibandingkan dengan tanpa menggunakan metode, 23,33% memiliki kecepatan yang sama dan 20% lebih lambat dari tanpa menggunakan metode.

4. KESIMPULAN

Berdasarkan penelitian Implementasi *Web Crawler* dengan *Ontology* yang diterapkan dalam studi kasus aplikasi pengarsipan didapatkan hasil sebagai berikut.

1. Pada proses penentuan objek relasi yang menjadi struktur *ontology* telah berhasil diimplementasikan dalam *web crawler* dengan presentase 100% objek relasi ontology sesuai dengan relasinya.

Pengindeksan *Web crawler* dengan *ontology* lebih cepat 56,67% dibandingkan pengindeksan tanpa *ontology*.

DAFTAR RUJUKAN

[1] Ahmat Josi, Leon Andretti Abdillah, Suryayusra.2014. Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah.

[2] Budi Yuwono, Savio L. Y. Lam, Jerry H.Ying, Dik L. Lee, 1996, *A World Wide Web Resource Discovery System*, in Proceedings of ICDE.

[3] Eri Zuliarso, Khabib Mustofa.2009. Crawling Web Berdasarkan Ontology.

[4] Subuki, Makyun. 2011. Semantik Pengantar Memahami Makna Bahasa. Jakarta : Transpustaka.

[5] Sukanta Sinha, Rana Dattagupta dan Debajyoti Mukhopadhyay.2003. Web-page Indexing based on the Prioritize Ontology Terms.

[6] Yusuf, Muhammad. "Apa itu Web Crawler". Muhammad Yusuf Gunadarma. 2012. Web. 23 Oktober 2016

[7] Zebua, Javier., 2010, Aplikasi Pencarian Buku Berbasis Web Semantik Untuk Perpustakaan SMK Yadika 7 Bogor , UniversitasGunadarma.