Information Retrieval System for Indonesian Manuscript using Semantic Web

Vihi Atina
Magister of Information System,
Diponegoro University

Eko Sediyono
Magister of Information System,
Satya Wacana Christian University

R. Rizal Isnanto
Magister of Information System,
Diponegoro University

ABSTRACT
The increase number of manuscripts and their diversity add the difficulty of searching and arranging for relevant manuscripts. The quality of search results provided by search engines has not been maximized in response to user requests because it does not involve semantic elements in the search process. It is necessary to build a information retrieval system for manuscript that makes it easier for researchers finding the title of the manuscript accordance with the topic of their research.

Information retrieval system for manuscript is built using semantic web. Manuscript data used in this research are Indonesian manuscript. Stages build system include data crawler process, build ontologies, NLP process, SPARQL query representation process and indexing process.

Information retrieval system for Indonesian manuscript can display the title and link of manuscript based on the search sentence entered. Tests are conducted on 3 types of search sentences with recall and precision methods. The recall value indicates that the owned manuscripts are returned 93.3% by information retrieval system. The precision value indicates that the results are returned 100% relevan by information retrieval system.

General Terms
Semantic Web

Keywords
Manuscript, crawler, semantic web, ontology, NLP, SPARQL query

1. INTRODUCTION
Increasing volume of information on web makes it hard to find, manage, access and maintain the information that is needed. There are at least two major causing that adversity. First, the meaning of information contained on web documents (web content), only can be understood by humans but can not be understood by the machine. As a result, the machine is unable to interpret what information is needed or sought by humans [1]. Second, now search engines are keyword-based search engines. These machines search for the documents based on the word (the spelling of the word) and not based on the meaning (the meaning of the word). This result causes irrelevant documents included as search results. It often happens that the relevant documents are not indexed by search engines. Human intervention is still needed to sort this information [2].

The semantic web pioneered by Tim Berners-Lee, is a way to represent web content in a form that can be understood and processed by an engine. Semantic web based on ontology is introduced by the W3C (World Wide Web Consortium) to bring significant progress in web search. Ontology becomes the cornerstone of many knowledge-based applications that require managing and data interpreting (usually text) from semantic perspective. Ontology is widely used to improve information retrieval [3]. In relation to search information on the web, ontology is useful to improve search accuracy. Ontology development represents the development process of the relationship between words or keywords model, so each word represents information [4].

The increase number of manuscripts and their diversity add the difficulty of searching and arranging for relevant manuscripts. The quality of search results provided by the search engines has not maximized in answering user requests because it is only based on word similarity and does not involve semantic elements in the search process. Ontology can be used for formal model in metadata semantic representation from manuscript and can be encoded using representation language. Manuscript can be easily arranged in searching process and makes reference of manuscript that interconnects.

Based on the background then it is necessary to build a information retrieval system for manuscript that makes it easier for researchers finding the title of the manuscript accordance with the topic of their research. Manuscript data used in this research are Indonesian manuscript. In this research can be formulated how to make information retrieval system for Indonesian manuscript using semantic web.

2. RELATED WORK
Two search techniques that have been developed are full text search and metadata search. Full text search is considered one of the most practical search methods in term operations since users simply enter a keyword, then the search engine will match this keyword to all available data. While metadata search means data from data, that is a collection of words organized with AND and OR logic. Each document is indexed and created for its metadata. The keyword search going to be matched with metadata that has been formed with certain restrictions, so the search process is simpler. Metadata search is better than full text search. However, from the discussion of the two search techniques has weakness that can not find synonym and homonym words [5].

Semantic search is understood as meaning search and solves concept limitations of keyword-based search engine. Semantic search improves search accuracy with understanding the search intent and contextual meaning as appeared in search data. Semantic search is search for content based on the proper context. Content is written text whereas context is condition for the existence of text. The purpose of semantic search is to look for content that fits with context. Semantic searching techniques can overcome weakness of full text and metadata methods [6].

Semantic search receives natural language query which is converted into semantic query. The query is matched to

semantic web database through semantic indexing technique. The result presented to user is semantic relevance sequence using dynamic ranking algorithm. This research can be upgraded to support multiple languages, to discover their interests, and to provide recommendation [7].

There are several models used for semantic search such as taxonomy, weighted tree similarity and ontology. Taxonomy is used to indicate the hierarchy of an object. Taxonomy has limitations because vocabulary is limited and inflexible. By using ontology to organize information, several words connected can be found, regardless of their hierarchical. The power of the word formed ontology is open and unlimited. When new information is formed, the information can be added to the ontology. While the weighted tree similarity semantic search model has structure that can not be shared then leads to waste in storage systems. This weakness can be overcome by ontology reusing the data structures that have been created [8].

Effective semantic ranking and search techniques deliver accurate results using ontology. The purpose of semantic search retrieves the highest grade results that are relevant with many significant query keywords. First, designing of weight measure determines the relative importance of semantic to determine the relevance of resources. Based on this measure, the ranking method calculates semantic relationships between resources and keywords as well as scope and strength of keywords discriminatory. The experimental results using data set show that the ranking method produces more accurate search than traditional method [9].

Some search engines have been able to identify language. Identification process is usually done by recognizing several words in document that are characteristic or specific for particular language, but the search engine does not analyze content of the document. As a result, for some search conditions it becomes very limited and even delivers results that have no connection with the meaning of the word that you want to find. One example of reliable search engine and the most widely used by users in Indonesia today is Indonesian Google (www.google.co.id). Indonesian Google has been chosen by many users because it has simple User Interface and can search in multiple URLs. However, Indonesian Google still has limitation especially analyzing document content in Indonesian language. Systems must be developed, so search engines understand Indonesian language with analyzing content of the text [10].

3. RESEARCH METHODOLOGY

3.1 Crawler

Crawler web is system that explores web hyperlink structure from an initial address and visits web address on a web page periodically. Search engine is one example of a large system that uses crawlers to traverse the internet constantly with aim finding and fetching as many web pages as possible. Here is the process crawler web [11]:

1. Download a web page.
2. Parse downloaded web page and retrieve all of links.
3. For each link taken, repeat the process.

3.2 Semantic Web

Semantic Web is defined as a set of technologies, which allows computer to understand the meaning of information based on metadata (information about the content of information). With metadata, computer is expected to interpret the results of information entry so that search results become more detailed and precise [12]. Semantic web technology is used to build systems by collecting content from different sources then to be processed, managed and shared to users. There are three important technologies involved in use of semantic web: eXtensible Markup Language (XML), Resource Description Framework (RDF), and Ontology Web Language (OWL).

3.3 Ontology

Ontology is theory about the meaning of an object, properties of an object, and relation of an object that may occur in domain of knowledge. Ontology development stage are described as follows [13]:

1. Determine domain and scope of ontology
2. Consider reuse existing ontology
3. Write down important words in ontology
4. Defines class and grade levels
5. Defines property
6. Defines role restriction
7. Create an instance (individual)

3.4 Natural Language Processing

Natural Language Processing (NLP) is application of computer science, especially computational linguistics, to examine interaction between computers with (natural) human language. NLP solves problems to understand natural language of human, with all its grammatical and semantic rules, and transforms language into formal representation that can be processed by computer [14]. NLP stage is case folding, tokenizing and filtering.

3.5 SPARQL Query

SPARQL is a protocol and query for data sources from semantic web. SPARQL executes queries instead of databases, but on data in RDF. To execute a SPARQL query, it need to know the resource, property and value of the RDF. The clauses used in the SPARQL query are *prefix*, *select*, *where* and *optional* [15].

3.6 Research Procedure

The research procedure is described in Information System Framework as shown in Fig 1.

Description of the information system framework as follows:

1. Document of Indonesian manuscript on web is done web crawler process
2. Data resulted from web crawler process are used as reference in building Indonesian manuscript ontology. Steps building ontology include:
    a. Specifies domain
    b. Specifies scope domain
    c. Create classes and their derivatives
    d. Create object properties
    e. Create data properties
    f. Create individuals
    g. Specifies restrictions for classes and properties
3. To display search results, User inputs keyword search, and then processes NLP. NLP includes case folding process, tokenizing, and filtering with reference stopword.
4. The next step, Query process is done to match the keyword and ontology, then indexing process is

done to represent search results into easily understandable form.

5. The search results is displayed in information retrieval system for Indonesian manuscript using semantic web. Display search results are list of titles and links Indonesian manuscript.

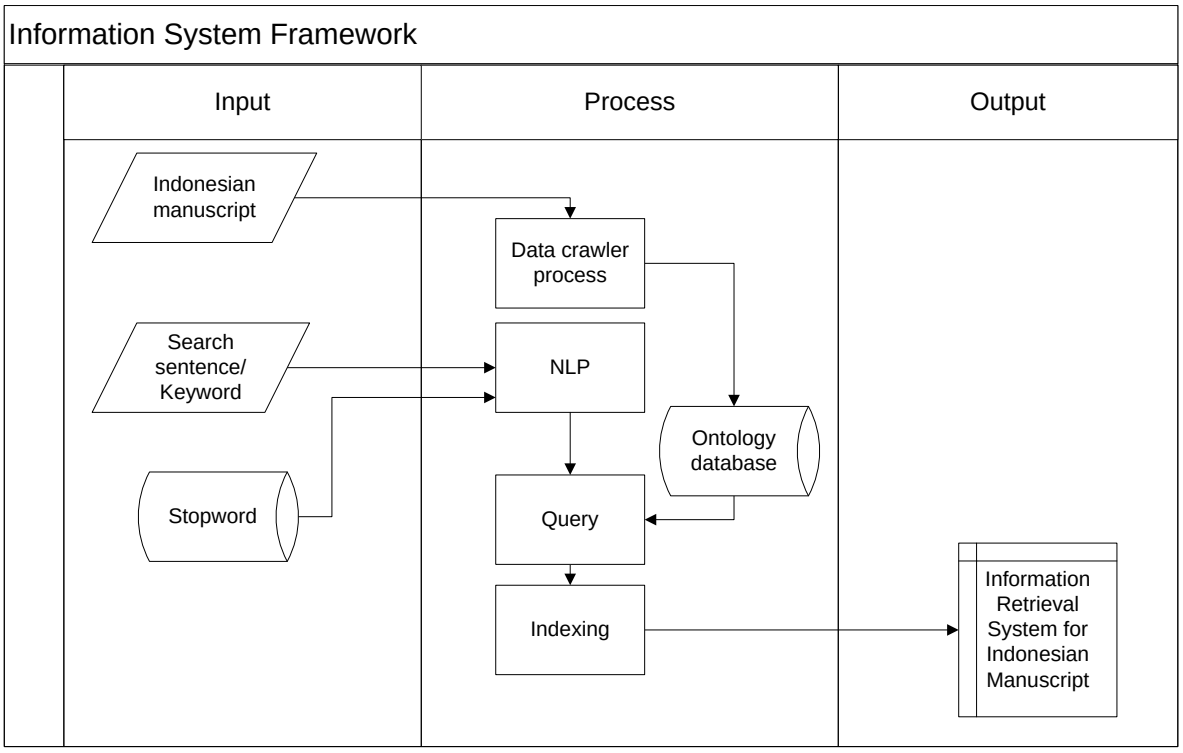

Fig 1: Information System Framework

4. EXPERIMENTAL RESULTS

4.1 Data Crawler Process

Data crawler process uses the Breadth-First Search algorithm. Concept of the Breadth-First algorithm checks every link in web page before it switchs to another web page. Breadth-First will browse every link on the first web page, then browses web page from every link the first web page, and so on. This search done until no new links are met . Breadth-First search tree can be seen in Fig 2.

Breadth-First Search algorithm in Fig. 2 consists of 4 levels starting from level 0, level 1, level 2 and level 3. Level 0 is the initial level searching manuscript on web address. Level 1 search all of journals and publishers that exist at level 0. Level 2 search all of journals edition that exist at level 1. Level 3 search all of manuscripts / papers, authors, abstracts and links that exist at level 2. The results of this process are become the basic data in building ontology. Example display results of data crawlers can be seen in Fig 3.

4.2 Build Ontology

Steps to build manuscript ontology as follows :

1. Create class
   Class consists of *jurnal* (journal), *penerbit* (publisher), *edisi* (edition), *naskah* (manuscript) and *penulis* (author).
2. Create object properties
   Object properties used in manuscript ontology are *terdiriDari* (consists of), *diterbitkanOleh* (published by) and *ditulisOleh* (written by). *terdiriDari* connect *jurnal* class and *edisi* class, as well as *edisi* class and *naskah* class. *diterbitkanOleh* connect *jurnal* class and *penerbit* class. *ditulisOleh* connect *naskah* class and *penulis* class.
3. Create data properties
   Data properties describe attributes of each ontology class. *Jurnal* class has attribute *namaJurnal* (name of journal), *penerbit* class has attribute *namaPenerbit* (name of publisher) , *edisi* class has attribute *namaEdisi* (name of edition), *naskah* class has attributes *judulNaskah* (title of manuscript), *abstrak* (abstract) and *linkNaskah* (link of manuscript), while *penulis* class has attribute *namaPenulis* (name of author).
4. Create individuals
   Individuals are made based on the results of data crawler process. Individuals consist of journals, publishers, editions, manuscripts and authors.
5. Create restrictions for classes
   penerbit class is equivalent to *publisher* class. *edisi* class is equivalent to *volume* class. *naskah* class is equivalent to *paper* class and *makalah* class. *penulis* class is equivalent to *peneliti* class and *pengarang* class.

The results of implementation manuscript ontology can be seen in Fig 4.

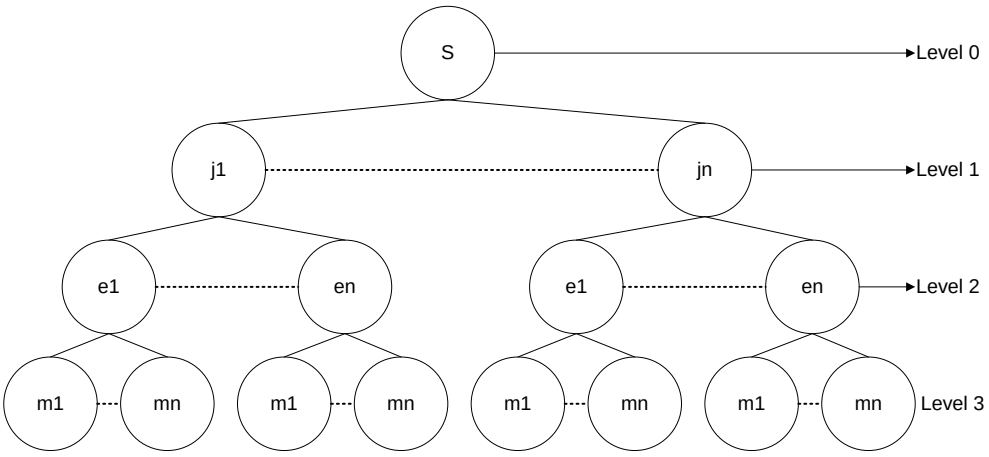

Fig 2: Breadth-First Search Tree

Perangkat Lunak Penganalisis Kemiripan Webpage Berdasarkan Konten Presentasional -> ?ref=browse& mod=viewarticle&article=358403
Perancangan Aplikasi Fuzzy Multi Criteria Decision Making (FMCDM) Untuk Menentukan Nilai Ketidakpastian Sistem Pakar -> ?ref=browse&mod=viewarticle&article=358404
Training Making Materials Video Interactive Learning For Teachers in SMK Negeri 1 Muara Enim -> ?ref=browse& mod=viewarticle&article=358405
Perbandingan Model Modifikasi Skema Pembiayaan Wired Internet Pada Jaringan Multi Kelas Multi Link Bottleneck -> ?ref=browse&mod=viewarticle&article=358406
Pengembangan Sistem Informasi Manajmen Data Sekolah Pada Dinas Pendidikan Kabupaten Ogan Komering Ulu -> ?ref=browse&mod=viewarticle&article=358407

Fig 3: Data Crawler Result

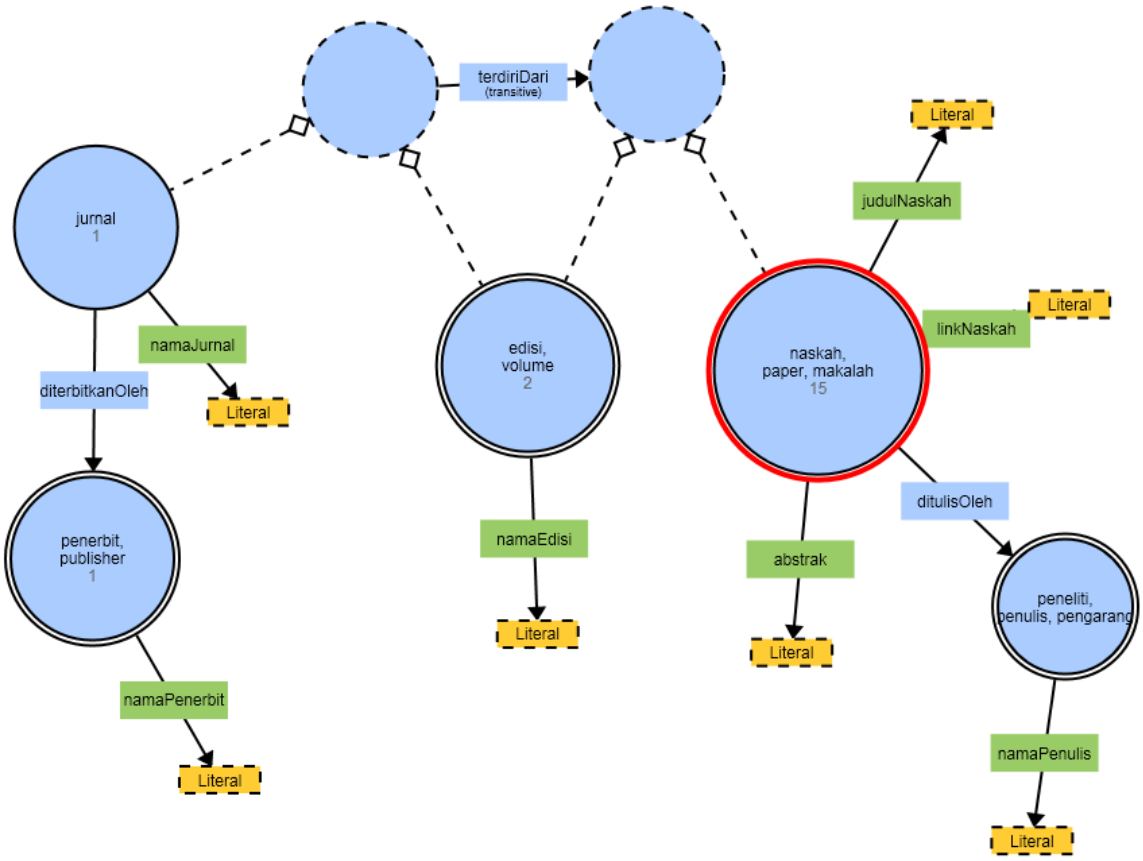

Fig. 4: Ontology Graph

4.3 NLP Process

Natural Language Processing (NLP) include case folding, tokenizing, and filtering. NLP process is performed after user input search sentence in information retrieval system.

1. Case Folding Process
   Process converts search sentence to lowercase. Example of case folding process can be seen in Fig 5.

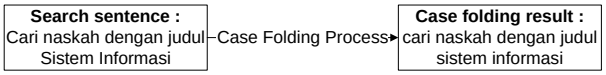

Fig 5: Case Folding Process

2. Tokenizing Process
   Process cuts string based on each word that feeds it. Sting used in this step is the result of case folding process. Example of tokenizing process can be seen in Fig 6.

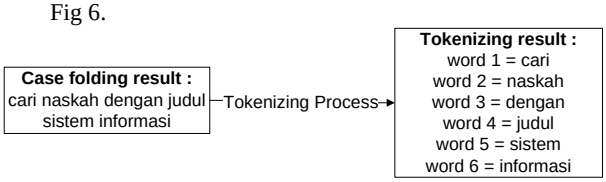

Fig 6: Tokenizing Process

3. Filtering Process
   Process takes important words from tokenizing result and discards the less important word (stopword). Example of filtering process can be seen in Fig 7.

Fig 7: Filtering Process

4.4 Query Representation

This step analyzes filtering result to determine type of search sentence. The analyzing results are used to interpret SPARQL query. Minimum sentence that has valid value for information retrieval system consists of: First, the minimum sentence should consist of command word (kp: search or display), and search categories (ctg: manuscript, paper) and value going to be searched. Second, the sentence which has datatype properties (dtp) or object properties (obp) more than one going to be declared valid if dtp or obp does not have the same meaning. There are 3 search sentence types.

1. Type 1 : K → kp + ctg + value
2. Type 2 : K → kp + ctg + dtp + value
3. Type 3 : K → kp + ctg +obp +value

Example of analyzing process can be seen in Fig 8.

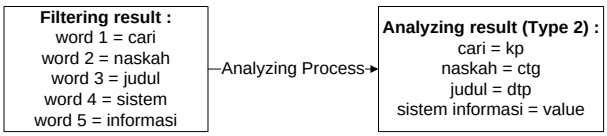

Fig 8: Analyzing Process

From Fig 8 show that analyzing result include in type 2, so SPARQL query reprentation as follows:

```
SELECT ?judulNaskah ?linkNaskah
WHERE { ?naskah :judulNaskah ?judulNaskah.
filter regex(str(?judulNaskah),'".$value."','i').
?naskah :linkNaskah ?linkNaskah}
```

4.5 Query Representation

Query represention result is matched with ontology data representation. Indexing result is list of titles and links manuscripts based on the search sentence entered. Example of information retrieval system result can be seen in Fig 9.

Fig 9: Information Retrieval System Result

5. EVALUATION

Testing is done by using recall and precision method. Recall and precision test is done using input search sentence from type 1, type 2 and type 3. Recall and precision calculation can be seen in Table 1.

Table 1. Recall and Precision Calculation

Type	Target	Selected	Relevan	Recall	Precision
1	2	2	2	100%	100%
2	5	4	4	80%	100%
3	3	3	3	100%	100%
Average				93.3%	100%

Test resullts from 3 types of search sentence show that recall value average is 93.3% and precision value average is 100%. The recall value indicates that the owned manuscripts are returned 93.3% by information retrieval system. The precision value indicates that the results are returned 100% relevan by information retrieval system.

6. CONCLUSIONS

Information retrieval system for manuscript is built using semantic web. Information retrieval system for Indonesian manuscript can display the title and link of manuscript based on the search sentence entered. Tests are conducted on 3 types of search sentences with recall and precision methods. The recall value indicates that the owned manuscripts are returned 93.3% by information retrieval system. The precision value indicates that the results are returned 100% relevan by information retrieval system. It is easier for researchers finding the title of the manuscript accordance with the topic of their research. This test is done with small-scale data, so to test the scalability system needs to be tried with larger-scale data. The results of this research can be re-used to develop information retrieval system for manuscript using semantic web with more complex ontology knowledge base.

7. REFERENCES

[1] Antoniou, G. and Harmelen, F. V. 2008. A Semantic Web Primer Second Edition. MIT Press. Cambridge.

[2] Lijun, T. 2011. The Study of Semantic Retrieval Based on The Ontologi of Teaching Management, Advanced in Control Engineering and Information Science. Procedia Enginering 15, 1555-1559.

[3] Castells, P., Fernández, M. and Vallet, D. 2007. An adaptation of the vector-space model for ontologi-based information retrieval. IEEETrans.Knowl.DataEng.19, 261–272.

[4] Kara, S., Alan, O., Sabuncu, O., Akpinar, S., Cicekli, N. K. and Alpaslan, F. N. 2012. An Ontologi-Based Retrievel System Using Semantic Indexing. Information System 37, 294-305.

[5] Jeffrey, B. 2008. The Weakness of Full-Text Searching. The Jaournal of Academic Librarianship, September 2008, 438-444.

[6] Faisal, R. 2009. Penerapan Algoritma Weighted Tree Similarity Untuk Pencarian Semantik Wikipedia. Master Thesis of Informatics Department. Surabaya.

[7] Thangaraj, M and Sujatha, G. 2014. An architectural design for effective information retrieval in semantic web. Expert Systems with Applications 41 (2014), 8225–8233.

[8] Jing, J. 2006. Similarity of Weighted Directed Acyclic Graph, New Brunswick : University of New Brunswik, Master Thesis..

[9] Lee, J. 2014. Effective ranking and search techniques for Web resources considering semantic relationships. Information Processing and Management 50 (2014), 132–155.

[10] Handayani, P. W., Wiryana, I. M. and Milde, J. 2008. Mesin Pencari Berbasiskan Semantik Bahasa Indonesia, Information System Journal MTI-UI, Volume 4, Number 2, ISBN 1412-8896.

[11] Shestakov, D., 2013. Intelligent Web Crawling. Department of Media Technology. Aslto University. Finland.

[12] Suteja, B. R. and Ashari, A. 2008. Ontologi e-Learning Content berbasis Web Semantik. SNATI 2008.Yogyakarta.

[13] Noy, N.F. and McGuinness, D.L. 2001, Ontology Development 101: A Guide to Creating Your First Ontology. Standford University.

[14] Pustejovsky, J. 2012. Natural Language Annotation for Machine Learning. O'Reilly. Beijing.

[15] Yadagiri, N. and Ramesh, P. 2013. Semantic Web and The Libraries: An Overview. International Journal of Library Science