Enhancing the Relevance of Searching Result of Songs' Information Using Ontology

Muhammad Ihsan JAMBAK[1*], Indah SURYANI[2] and Dedy KURNIAWAN[1]

[1] *Informatics Management Department, Sriwijaya University, Palembang, Indonesia*
2 *Informatics Department, Sriwijaya University, Inderalaya, Indonesia*
Corresponding author: jambak@unsri.ac.id

ABSTRACT

The crawler and Machine Learning based search engine are not optimal in determining relevant search results. The aim of this research, therefore, is to increase the relevance of search engine results. This design was carried out on a semantic web, using ontology with RDF due to its knowledge representation, and SPARQL as a query to retrieve information contained in the RDF Ontology. In this study, the song information ontology framework was designed and implemented using the Protégé tool with the SPARQL query language. Mapping in SPARQL queries affects the speed and relevance of the application while displaying search results. The results of the design were tested using simulated web applications with the Java Server Pages (JSP) programming language and search techniques were categorized. Information used as the object is obtained from songs which deals with the data at the knowledge level using semantic technologies delivered from steady and trustworthy base. An initial evaluation of the system showed favorable results from the tests conducted on 100 sample songs, with 100% relevance and precision ratio of 1:1.
Keywords: ontology, semantic web, XML, RDF

INTRODUCTION

Fast and accurate information is believed to be one of the primary needs of human, with data digitally stored due to the presence of varying numbers. The internet allows access to millions of web pages, thereby, making it easier for the extraction of various needed information. An interesting issue centered on how to semantically store information for easy reusability.

However, the creation of search engines helped to facilitate information retrieval from the internet. This web application consists of two essential components namely crawlers and machine learning. The primary function of the Crawler is to browse and retrieve web pages, with the results indexed by the search engines to facilitate information processing [1]. While, machine learning is a means of obtaining relevant documents by filtering irrelevant data on the Web (URLs) [2]. Initially the system used to extract information from search engines, were not knowledge-based and not semantically performed, thereby, leading to irrelevant results. This led to the development of the ontology model approach, which supports a Knowledge Management System and opened the possibility for web pages to move from a document-oriented view to an interrelated knowledge that is flexible and dynamic. Ontology depicts a means of representing knowledge such as the meaning, property, and relationship of objects in the domain of knowledge [3-6]. Semantic webs have provided new strategies to increase the effectiveness of documents searches on the internet or specific electronic data sources. The intelligent search engine is expected to help internet users to conduct necessary and relevant searches.

SEMANTIC WEB

Semantic web is a web-based data which is directly or indirectly processed by inputting data in a form that is easily interpreted by the computer [7]. The aim of the semantic web is to develop standards and technologies that aid computers in understanding information on the Web. This helps to support productive browsing, data integration, navigation, and automation of tasks. Furthermore, this method helps in the delivery and utilization of information on the World Wide Web (WWW).

The semantic web method converts HTML-based data into a format understandable by the machine, this aids the computer in gathering, understanding and processing information. For example, with the current web technology, when people need to compile information from several websites at once, they need to visit these sites individually, to copy and paste content in order to create comprehensive information. This is a waste of time and energy because was manually conducted. However, the HTML-based web pages were designed to be understood by humans rather than machines [8].

The semantic web is designed in such a way that the computer is able to interpret information based on metadata (contents). With the existence of metadata, computers are able to automatically interpret and process information in order to deliver accurate results. Raw data is processed using semantic technologies delivered by a steady and trustworthy base [8, 9].

The most important standards in building a semantic web are XML, XML Schema, RDF, OWL, and SPARQL [10, 11]. It allows web developers to add one layer "meaning" to their documents, and aids the computer to scan and interpret information using software agents that crawl all over the web. This is possible because the semantic web has gathered information on what is called ontology [7]. The layers on the semantic web as recommended by the World Wide Web Consortium (W3C), is further explained as follows [12-15]:

Internationalized Resource Identifier (URI) is a unique name for semantic web identification.

Unicode presents and manipulates text into many languages.

XML is a language that allows programmers to create web documents based on their vocabulary. It is suitable for sending documents through the web.

RDF is a data model for objects (resources) and the relationships between them. It provides a simple meaning for the data model, which is represented in the XML syntax.

RDF Schema provides the basics of vocabulary and allows for the possible creation of classes and property hierarchy.

The ontology of vocabulary extends to the RDF Schema by adding more sophisticated concepts, such as a cardinality, boundary values, and transitive properties. This concept is based on logic to provide the power of reasoning on the semantic web.

Logic is used to further improve the language of ontology and allows writing from a declaration of knowledge.

The proof involves a subtraction process as well as representing evidences in web languages and proof validation.

Trust aims to ensure and verify that semantic web statements come from trusted sources. This trust is precisely achieved using 'digital signature' from the RDF statement.

ONTOLOGY

Ontology is a file which defines the relationship between terms and groups. It is considered one of the pillars of the semantic web usually accompanied by documents in formal language. Ontology need to be detailed, comprehensive, and effective for the semantic web to work. In addition, it is widely used to primarily support semantic web, such as technology designed to understand the meaning of a word or sentence entered by the user. Designing computers to function like the human brain is something that seems impossible, but this vision is continually pursued by providing a set of tools that aids machines or computers to easily process information and understand data entered by the users [5, 6, 11].

Ontologies consist of formal or defined language structures which include the following XML (Extensible Markup Language), RDF (Resource Description Framework), RDF Schema, and OWL (Ontology Web Language) [16, 17]. Technically ontology is represented in form of classes, properties, slots, and instances.

The class explains the concept (or meaning) of a domain. It is a collection of elements with the same property, and consists of subclass derivatives which describe more specific concepts.

Property explains the concept of measurable values and statuses which may exist within a domain.

The slot is a representation of the knowledge framework or relation that explains the properties of the class and instance.

The instance is the individual derived from a subclass. It is similar to the instance from a superclass.

EXTENSIBLE MARKUP LANGUAGE (XML)

XML is a universal meta-language for defining markup. It provides a general framework and designed to create an easy means of sending data and metadata via the Web [12]. Unlike HTML, it allows users to define needed custom tags [8]. However, the standard XML does not have semantic limitations on the meaning of any document. Its elements represent codes written and compiled in documents with opening tags, contents, and closing tags.

Example of XML tags:
<lecturer>

DavidBillington

</ lecturer>.

XML Schema is a language used to define a set of rules that XML documents need to adhere. It allows a user to establish or replace a new/existing type. A combination of XML syntax features build rules and reduces workload.

RESOURCE DESCRIPTION FRAMEWORK (RDF)

RDF is a metadata used to describe the resource address on the web in the form of titles, authors, copyrights, and licenses. Statements in RDF consist of subjects, predicates, and objects. The subject is any described term usually called a URI address which replaces a resource. Predicates are defined as properties of resources which explain the relationship between the subjects and objects. The object consists of two data types URI such as http://musikday.com/lagu/genre and literal eg "adam air." Subjects and predicates contain data made of resources, while objects are either resource type or literal [18].

RDF uses a graph to represent a collection of statements with vertices depicting entity, and arrows representing the relationships between objects. Triples are called "RDF graphs," with each constituting of a statement that describes the relationship between things denoted and addressed by the node [19]. It is designed to represent minimal and flexible information, with separate applications individually designed in a direct and understandable manner. Furthermore, it generally provides higher added value because it is shareable with increase in information value due to its accessibility on the internet.

RDF is a universal language that permits users to explain resources using their own vocabulary. However, it does not have the tendency to assume of define the meaning of domains. Therefore, its scheme which serves as a data dictionary for describing properties and resources of its was designed to take care of this deficiency [8].The mechanism adopted by RDF to regulate expressions on boundaries is to make resources, properties, types, and statements the main objects on the web. The expression has a URI which is not limited to the basic level to be combined.

ONTOLOGY WEB LANGUAGE (OWL)

OWL was designed by the W3C ontology web working group and is intended to be the successor of DAML + OIL. It is the most expressive representation of knowledge for the semantic web [20-22]. OWL was designed to be used by applications that need to process information to users. It facilitates more significant interpretation of web content, which is supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary and formal semantics.

In 2004, the W3C officially released the OWL as a recommendation language to present ontology. It was developed from the description of logic and DAM + OIL. This language is used by applications which display information to users and programmers [8]. OWL tends to add new vocabulary in addition to formal semantics as made before XML, RDF, and RDF Schema [10]. This OWL is very effective in interpreting web content on the machine and offers three sub-languages with different levels for the specific needs of users, namely OWL Lite, OWL DL, and OWL Full.

SPARQL

Mapping in a SPARQL query affects how quick the application displays its search results with the provision of a standard query language for RDF charts. Similar to SQL and XQuery, the SPARQL language offers a declarative interface for interacting with RDF databases. It is a query language designed to collect data from various sources and accelerate the development of web 2.0 applications, thereby creating a standard service for inquiries [23-25].

SPARQL is run using several tools and APIs such as ARQ, Rasqal, RDF Query, twingql, Pellet, and KAON2. These tools consist of APIs that make programmers manipulate query results with various applications. However, the SPARQL Query Results XML Format recommended by the W3C is usually used as a standard [10].

Research Methods

In this study, an ontology-based web application was built in order to discover information relating to songs. Secondary data was used from several song provider sites with the title, singer, release, url, lyrics, genre, and feelings manually obtained by copying the URL in order to obtain the address of the * .mp3 format of the song and detailed information from the webpage. A total of 100 songs were taken as samples from five music genres in Indonesia, namely Pop, Rock, Jazz, Dangdut, and Reggae. The data collected was categorized as taxonomy and processed based on the ontology effectiveness.

Ontology Design

Ontology which is also referred to as Informasilagu, is present in the design of all semantic-based applications. Its structure is as follows:

a. Lagu is the class used to define the concept of all information obtained from the song. This class is a domain which consists of two property objects, namely hasGenre and hasPerasaan, used to link classes. This class consists of five data properties which serve as criteria for the knowledge base. It consists of judul (title), penyanyi (singer), rilis (release), lirik (lyric), and url.

b. Genre is the class which ranges from hasGenre. It is used to define the type of music and its concept. The genre of a song has one data property, called jenis (type).

c. Perasaan, is any class used to describe the feelings of the song and it has only one data property known as suasanhati.
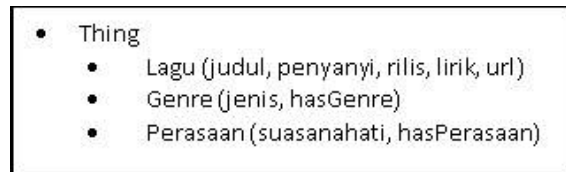

Figure 1: Ontology Structure of Informasilagu

After the class and data property were defined, it was discovered that a link or form of relationship was needed to form a connection. Class acts as an ontology subject, while the data property is an object. The predicate in ontology is defined as the relationship between a subject and an object in order to have uniform meaning as shown in table 1.

Table 1. Domain-Range Ontology of Informasilagu

Domain	Object Property	Range
Lagu	hasGenre	Genre
Lagu	hasPerasaan	Perasaan

An ontology scheme designed with Protege software is shown in figure 2. In the class of Lagu, each song has some identical information such as title, singer, release date, lyrics, and url. The songs also have a property object which either belongs to hasGenre or hasPerasaan to show its feeling to the listener. Genre and Perasaan have two data properties, namely jenis (type) and suasanahati (feeling)
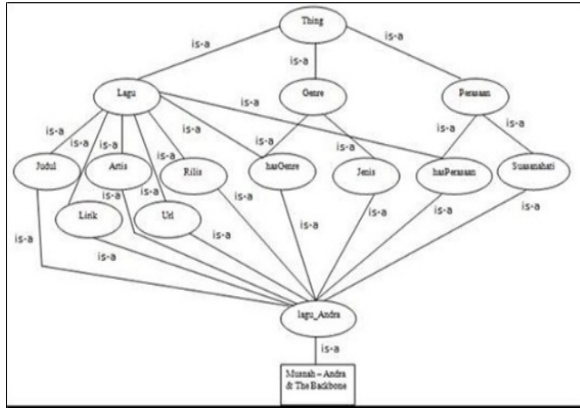

Figure 2: Ontology Scheme of Informasilagu

RDF DESIGN

The RDF design was then developed to explain the relationship between resources and literal, which shows the position of the subject, predicate, and object. RDF is visualized in the form of a graph which explains its flow. The resources are illustrated in oval shaped boxes while the literal form as explained in figure 3. The figure is a data model for Indonesian song search ontology which describes the relationship between property and domain. This is illustrated in table 2.
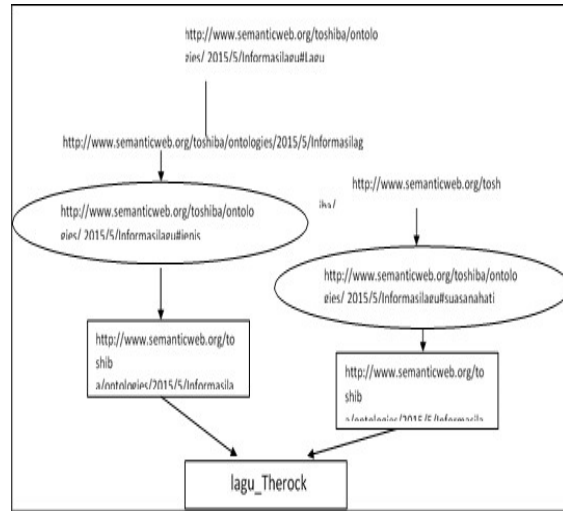

Figure 3. RDF Graph

Table 2. Ontology Property Description

Property Name	Domain	Range	Property Type
judul	Lagu	string	Data Property
penyanyi	Lagu	string	Data Property
lirik	Lagu	string	Data Property
rilis	Lagu	string	Data Property
url	Lagu	string	Data Property
jenis	Genre	string	Data Property
suasanahati	Perasaan	string	Data Property
hasGenre	Lagu	Genre	Object Property
hasPerasaan	Lagu	Perasaan	Object Property

ONTOLOGY INSTANCE

The instance simply means inputting data into an ontology using the Protégé software editor. The sample used in this study was 100 Indonesian songs, with instance process conducted with raw data entered into the various fields. Table 3 shows an example of how to design instance ontology.

Table 3. Instance Design

Type	Instance
Lagu	Lagu Andra judul Musnah, Penyanyi Andra and The Backbone, rilis, lirik, dan url
Genre	Genre Rock Jenis Rock
Perasaan	Perasaan_Sakithati suasanahati Sakit Hati

QUERY DESIGN

The language for ontology queries is SPARQL, which is an acronym for Simple Protocol and RDF Query Language. The SPARQL standard defines network protocols used to exchange queries and languages. It adopts an SQL-like syntax to show queries, with web data represented using RDF data format and specific query language. To obtain information from an RDF graph a query SPARQL language is needed to send and receive results.

```
PREFIX Informasilagu:
<http://www.semanticweb.org/toshiba/ontologies/2015/5/Informasilagu#>
SELECT ?instance  ?judul
WHERE {
    ?instance rdf:type Informasilagu:Lagu .
    ?instance Informasilagu:judul ?judul
}
```

```
PREFIX Informasilagu:
<http://www.semanticweb.org/toshiba/ontologies/2015/5/Informasilagu#>
SELECT ?instance  ?judul ?penyanyi ?rilis ?url ?lirik
WHERE {
    ?instance rdf:type Informasilagu:Lagu .
    ?instance Informasilagu:judul ?judul .
    ?instance Informasilagu:penyanyi ?penyanyi .
    ?instance Informasilagu:rilis ?rilis .
    ?instance Informasilagu:url ?url .
    ?instance Informasilagu:lirik ?lirik .
}
```

Figure 5. The Query design displays all Instances from the domain of Lagu

RESULTS AND ANALYSIS

The first testing was conducted using several search queries that were inputted according to the needed category, namely title, singer, lyrics, genre, or taste. This makes it easier for the search query to be processed by the system in order to locate the needed song. The results show that the system successfully processed the inputted keywords, which are stored in ontology (OWL), in the form of judul (title), penyanyi (singer), lirik (lyrics), jenis (type), and perasaan (feeling). However, results are obtained despite the number of word/s query as long as it is correctly spelt.

Table 4. Query Based Search Test Results

No.	Query	Data Property	Output	Number of File	Result
1	Musnah	judul	Musnah	1	Accepted
2	afgan	penyanyi	Afgan Rossa ft. Afgan	2	Accepted
3	Aku tanpa langkah kakiku	lirik	Lirik lagu Noah – Seperti Kemarin	1	Accepted

4	Dangdut	jenis	Muara Kasih Bunda-Erie Susan Klepek klepek-Hesty Pusing Pala berbie- Putri Bahar Nikahin Aku- Bebizie Aku Mah Apa Atuh-Cita Citata Masa Lalu-Inul Daratista Cabe cabean-Via Vallen Goyang Dumang-Cita citata Ratu Hatiku-Saiful Jamil	9	Accepted
5	Sakit hati	suasanahati	Seperti Kemarin – noah Sandiwara Cinta-Repvblik Penipu Hati – Tata Janeeta Jangan Kau Bohong- Fatin Aku Mah Apa Atuh-cita citata Masa Lalu- Inul Puing Kenangan-Ungu Baru –Tulus Sakit Hati- Yovie and Nuno Musnah – Andra and the Backbone	10	Accepted
6	Kita	judul	Rumah Kita – God Bless Di Antara Kita - Adera	2	Accepted

The second testing was based on Noname Song in Ontology, with keywords that were not in the property data in the ontology (OWL) file used as search queries. However, there might be a problem finding results assuming the query does not match the created ontology file. This is shown in table 5.

Table 5. Test Results Based on Noname Songs in Ontology

No.	Query	Data Property					Number of songs	Result
		judul	penyanyi	lirik	genre	perasaan		
1	Percayalah	-	-	-	-	-	-	-
2	Musnah Mp3	-	-	-	-	-	-	-
3	Sperti Kemarin	-	-	-	-	-	-	-

The third testing was based on Crawler and Machine Learning. In this study, stafaband.co was used as a contrast website because it contained same contents. Some results of the test are good as shown in Table 6, however, there are certain irrelevant answers. For example, the search for the song "sabar," produced a song titled "Bersabar" and "Yang Kusebut Sayang" from an album "Cinta ... Bersabarlah". These results prove that the web with crawlers and machine learning needs to be improved.

Table 6. Crawler & Machine Learning Based Web Results

No.	Query	Data Property	Output	Album	Result
1	afgan	penyanyi	Afgan feat Raisa – Percayalah Afgan – Knock Me Out Afgan – Ku Mohon	Single Pop 2015 Single Maret 2015 Single Religi 2014	Accepted Accepted Accepted
2	sabar	judul	Saykoji – Bersabar Fatin – Sadar Dibatas Sabar Letto – Cinta … Bersabarlah Letto – Yang Kusebut Sayang Afgan - Sabar	Single Sept. 2015 For You 2010 Cinta … Bersabarlah Cinta … Bersabarlah Live to Love	Rejected Accepted Rejected Rejected Accepted
3	aku yang paling kau cinta	lirik	Mojo – Roman Cinta Berlian Band – Si Ratu Tega Bersamamu - Putih		Rejected Rejected Rejected

The fourth testing was searching for song information using ontology. This was tested using 100 song files by implementing Simple Search, which a search facility is based on the categories provided. Simple Search was chosen because the results were more accurate than the keyword search [26]. Here testing is conducted individually in all categories in order to ascertain the accuracy of the results.

Table 7 - Test Results for Song Searching with Ontology

No.	Category	Number of Files	Result
1	Judul	100	Found
2	Penyanyi	100	Found
3	Lirik	100	Found
4	Genre	100	Found
5	Perasaan	100	Found

Effectiveness and efficiency were measured by studying the values of Recall, Precision, and F-Measure. Precision is the number of relevant documents obtained from search results divided by the total number of search results. The recall is the number of relevant documents found by the system divided by the number of relevant documents available. F-Measure is used to calculate effectiveness. The results obtained from testing this system is show in Table 8.

Table 8 – Comparison of Ontology-Based vs. Crawler and Machine Learning Based

No	Query	No. of Files	Crawler & Machine Learning Based						Ontology-Based					
			A	B	C	R	P	F-M	A	B	C	R	P	F-M
1	Judul	100	86	0	14	100%	86%	92,47%	100	0	0	100%	100%	100%
2	Penyanyi	100	100	0	0	100%	100%	100%	100	0	0	100%	100%	100%
3	Lirik	100	74	11	15	87,05%	83,14%	85,05%	100	0	0	100%	100%	100%
4	Genre	100	100	0	0	100%	100%	100%	100	0	0	100%	100%	100%
5	Perasaan	100	-	-	-	-	-	-	100	0	0	100%	100%	100%
			Average			96,76%	92,28%	94,38%	Average			100%	100%	100%

A : Relevant items retrieved   R   : Recall (A / (A + B) * 100%)
B : Relevant items not retrieved   P   : Precision (A / (A + C) * 100%)
C : Irrelevant items retrieved   F-M   : F-Measure (2 * P * R) / (P + R)

CONCLUSIONS

Numerous information are found on the internet, therefore, search mechanism such as the Information Retrieval System is required to find relevant information based on the data entered by the user. According to the results of the Crawlers and Machine learning based, the average value of Recall is 96.76%, Precision is 92.28%, and F-Measure is 94.38%. These results prove that the system on this web is excellent, however, assuming ontology is used, the relevance increases and the values of Recall, Precision, and F-measure reaches 100%. These results prove that the ontology-based search engine system is better with a 1:1 ratio for Recall and Precision, which means that this application has high effectiveness and efficiency. The aim of building this system is to determine the keywords with SPARQL Query. However, this application is also able to understand and provide relevant results for various words searched.

REFERENCES

[1]   E. Zuliarso and K. Mustofa, "Crawling Web berdasarkan Ontology," Dinamik, vol. 14, 2009.

[2]   M. Chau and H. Chen, "A machine learning approach to web page filtering using content and structure analysis," Decision Support Systems, vol. 44, pp. 482-494, 2008.

[3]   S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," Computer networks, vol. 31, pp. 1623-1640, 1999.

[4]   A. Thomasson, "The ontology of art and knowledge in aesthetics," The Journal of Aesthetics and Art Criticism, vol. 63, 2018.

[5]   R. Arp, B. Smith, and A. D. Spear, Building ontologies with basic formal ontology: Mit Press, 2015.

[6]   T. Gruber, Ontology: Springer, 2009.

[7]   T. Berners-Lee, M. Fischetti, and M. L. Dertouzos, Weaving the web-the original design and ultimate destiny of the World Wide Web by its inventor. San Francisco, CA: Harper Business 2000.

[8]   M. Awaludin, "Sistem Navigasi dan Pencarian Berbasis Konteks Pada Konten E-Learning Menggunakan Teknologi Web Semantik," Institut Teknologi Sepuluh November. Surabaya, 2009.

[9]   M. Á. Rodríguez-García, L. O. Colombo-Mendoza, R. Valencia-García, A. A. Lopez-Lorca, and G. Beydoun, "Ontology-based music recommender system," in Distributed Computing and Artificial Intelligence, 12th International Conference, 2015, pp. 39-46.

[10]   N. Ibrahim, "Pengembangan aplikasi Semantic Web untuk membangun web yang lebih cerdas," Jurnal Informatika, vol. 3, pp. 27-39, 2007.

[11]   D. L. McGuinness and F. Van Harmelen, "OWL web ontology language overview," W3C recommendation, vol. 10, p. 2004, 2004.

[12]   G. Antoniou and F. van Harmelen, "A Semantic Web Primer" second," ed: The MIT Press Cambridge, Massachusetts, London, England, 2008.

[13]   Y.-F. Li, J. Z. Pan, S. Krishnaswamy, M. Hauswirth, and H. H. Nguyen, "The ubiquitous semantic web: Promises, progress and challenges," in Web-Based Services: Concepts, Methodologies, Tools, and Applications, ed: IGI Global, 2016, pp. 272-289.

[14]   R. Volz, S. Staab, and B. Motik, "Incrementally Maintaining Materializations of Ontologies Stored in Logic Databases," in Journal on Data Semantics II, S. Spaccapietra, E. Bertino, S. Jajodia, R. King, D. McLeod, M. E. Orlowska, et al., Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1-34.

[15]   D. Fensel, K. P. Sycara, and J. Mylopoulos, "The Semantic Web-ISWC 2003," in Second International Semantic Web Conference, Sanibel Island, FL, USA, 2003.

[16]   E. Hirsch, "Language, ontology, and structure," Noûs, vol. 42, pp. 509-528, 2008.

[17]   I. Horrocks, P. F. Patel-Schneider, and F. Van Harmelen, "From SHIQ and RDF to OWL: The making of a web ontology language," Web semantics: science, services and agents on the World Wide Web, vol. 1, pp. 7-26, 2003.

[18]   I. W. S. Wicaksana and W. Simri, "Ontology: Bahasa dan Tools Protégé," Universitas Gunadarma, Indonesia, 2006.

[19]   L. Provoost and E. Bornier, "Service-oriented architecture and the semantic web: a killer combination," Department of Information and Computing Sciences University of Utrecht, 2006.

[20]   M. C. Daconta, K. T. Smith, and L. J. Oerst, "The Semantic Web: a guide to the future of XML, Web services, and knowledge management," Computing Reviews, vol. 45, pp. 778-779, 2004.

[21]   M. Ahmad, "An ontology-based approach for integrating engineering workflows for industrial assembly automation systems," University of Warwick, 2017.

[22]   D. Campbell, "An Ontology-Driven Approach To Personalised mHealth Application Development," Edge Hill University, 2018.

[23]   T. Kanimozhi and A. Christy, "Applications of Ontology and Semantic Web in Image Retrieval and Research Issues."

[24]   T. Segaran, C. Evans, and J. Taylor, Programming the Semantic Web: Build Flexible Applications with Graph Data: " O'Reilly Media, Inc.", 2009.

[25]   H. Wijayanto, W. L. Y. Saptomo, and T. Susyanto, "PENERAPAN WEB SEMANTIK DALAM PENCARIAN KATALOG BUKU DI PERPUSTAKAAN STMIK SINAR NUSANTARA SURAKARTA," Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN), vol. 1, 2013.