INDONESIAN-ENGLISH CROSS-LINGUAL LEGAL ONTOLOGY FOR INFORMATION RETRIEVAL

Eri Zuliarso [1] , Retantyo Wardoyo[2],Sri Hartati[2],Khabib Mustofa[2]

[1] Student of Computer Sciences Post Graduate Program, Gadjah Mada University
[2]Department of Computer Sciences and Electronics Instrumentations,
Gadjah Mada University, Indonesia

ABSTRACT

This research encompasses the construction of a multilingual lexical database for cross-lingual information retrieval in the Indonesian legal domain. Multilingual lexical database featuring lexically and legally grounded conceptual representation can fit the cross-lingual information retrieval. Lexical database use Ontology Web Language (OWL) representation language. This representation is useful to provide application developers a high-quality resource and to promote interoperability.

KEYWORDS

Lexical Database , Cross-lingual information retrieval, Legal Domain, Ontology Web Language (OWL)

1. INTRODUCTION

Legal information is one of important information for personal development and social environment. Also, it is very important to the fulfillment of human rights and constitutional rights of citizens. The need for compliance with the legal information is based on the principle of law that a rule at the time was passed, immediately have binding legal force. Ignorance of the law can not be an excuse[16]. Thus apply the legal fiction that everyone knows all the rules of law. The legal fiction in fact indirectly provide obligation fulfillment of the right to information law. If the right to information law is not met, then the legal fiction that will create injustice. Based on the above reasoning, information law should be positioned as public property. Information law is a constitutional right of every citizen. The state, which in this case is executed by all state officials must fulfill that right without discrimination.

Legal databases are syntactically structured text archives with powerful search engines. But, search engines for legal information retrieval do not include legal knowledge into their search strategies. These strategies include keyword and metadata search, but do not address the semantics of the keywords, which would allow, for instance, conceptual query expansion. In other words, there is no semantic relationship between information needs of the user and the information content of documents [13].

A legal ‘language’ consisting of a complex structure of concepts, forms an abstraction from the text corpus as represented in legal databases[7]. Such legal structural knowledge does not only contain interpretations of the meaning of legal terms, but also shows logical and conceptual structure. Bridging the gap between legal text archives and legal structural knowledge is the key challenge in legal information retrieval[14].

DOI : 10.5121/ijwest.2015.6401

This is especially a problem for legal cross-lingual information retrieval. In this case, lack of knowledge of a certain language may prevent users from formulating queries and finding relevant results [14].

For several years, legal ontologies have been developed in a variety of projects that have a concern in the development of legal knowledge and law information management. Some of them are LRICore [4] and Jur-IWN [5].

LRI-Core used by the Dutch criminal law ontology of the e-Court project to support knowledge acquisition [3]. Also, the idea was to ground or anchor the concepts of these criminal law domain ontology regarding Dutch law to LRI-Core, to ease the process of construction of other criminal law domain ontologies (Polish law and Italian law).

Jur-IWN is an extension of ontology-based legal domain of the Italian version EuroWordnet. In Jur-IWN, synset associated with some semantic relationships such as hyperonim, hyponymy, hypernymy, meronym or instance-of. Jur-IWN ontology provides a lexical database that supports information retrieval systems and facilitates access to multilingual data [9].

Modelling knowledge by using ontologies or advanced thesauri enhances the ability to extract and exploit information from documents [7]. This is done by establishing explicit semantic links among related items. An ontology is an explicit formal specification of a common conceptualisation [10]. A formal definition of term hierarchies, relations and attributes (the explicit description of concepts in the legal domain) opens the way for implementations, such as information retrieval systems.

2. INDONESIAN LEGAL CONCEPT

Article 1 point 2 Law Of The Republic Of Indonesia No. 12 Of 2011 about Concerning Making Rules stated that the definition of Rules are written regulation that contain legal norms binding in general and formed or determined by a state agency or official authorized by the procedures specified in the Rules. From these definitions, the elements of the Rules are are written regulation, formed or determined by a state agency or official authorized, and contain legal norms binding.

Type and hierarchy of rules in Indonesia according to article 7 paragraph 1 Law Of The Republic Of Indonesia No. 12 of 2011 consists of:

- Constitution of the Republic of Indonesia of 1945,
- People's Consultative Council Decree, Law / Government Regulation In Lieu of Law,
- Government Regulation,
- Presidential Regulation,
- Province Regulation,
- Regency / Municipality Regulation.

"Hierarchy" is the level of each type of Rules based on the principle that the lower Regulations must not conflict with a higher Regulations.

"Framework" according to the Oxford Dictionary [18] can mean the structure of a particular system. Framework of Rule according Attachment I Law Of The Republic Of Indonesia No. 12 Of 2011 consists of:

A. TITLE
B. OPENING
1. By The Grace Of Almighty God phrase
2. The Former Position Of Rule Maker
3. Consideration
4. Legal Basis
5. Dictum
C. BODY
1. General Provisions
2. Basic substances are regulated
3. Criminal provisions (if required)
4. Transitional provisions (if required)
5. Closing Provision
D. CLOSING
E. EXPLANATION (if required)
F. ATTACHMENT (if required)

2.1. Legal basis

The legal basis beginning with "In view of" or "Referring to:" . The legal basis includes:

a. The basis authority to make Rule; and
b. Rules that order the establishment of the Regulation.

Regulations are used as the basis of the Regulations only same level or higher Regulations. The order inclusion of a legal basis needs to consider hierarchy of rules, if the legal basis are more than one. The legal basis arranged chronologically according to when the enactment or stipulation if the levels are same.

2.2. General Provisions

General provisions laid out in chapter one. If the regulation does not do the grouping chapter, the general provisions laid out in a chapter or a few chapters early. General provisions may contain more than one chapter. Some specific terms used in the regulation defined in the General Conditions section.

In attachment explain that General provisions contain:

a. limit of understanding or definition;
b. abbreviation or acronym as outlined within the limits of understanding or definition; and/or
c. other matters of a general nature applicable to the article or a subsequent article include provisions that reflect the principles, purposes, and objectives without separately formulated in the article or chapter.

The limitations of meanings or definitions, abbreviations, or acronyms serve to explain the meaning of word, therefore it must be formulated with a complete and clear so as to avoid double meaning. Word contained in the general provision is simply word that is used repeatedly in a chapter or a few chapters later. Nevertheless, the word is defined though is only used once. This is due to the word that required understanding for a chapter, section or paragraph.

Formulation of limit the understanding of the Rule may vary with the formulation of other Rule. This variation occurs because of adaptation as needed with the substance to be regulated. Implementation regulation is sometimes necessary to quote higher regulations. Under these

conditions, the formulation of meaning or definition in the implementing regulation must be equal to the formulation of meaning or definition contained in the meaning or definition of the higher regulation is implemented.

The General provisions (limit of understanding or definition) are useful that can be used as glosses for the corresponding word to building the lexical database. As to second definition techniques, abbreviation can be used as synonyms for words.

In a context of cross-lingual information retrieval, the links between words in different languages have to be established on the basis of their meaning. A parallel document used to establish the link between term and definition in both Indonesian and English.

3. Building Lexical Database

Lexical database creation process described in flowchart in Figure 1. Making lexical database initiated by crawling regulation documents on the website http://traderulebook.ekon.go.id/. This website maintained by Coordinating Ministry For Economic Affairs Republic Of Indonesia. Then, words and their definitions get from extraction in the general provisions. After getting the words and their definitions, link analysis is performed to verify consistency of word definitions in a regulation with word definitions in other regulations. Link analysis has been conducted by examining the legal basis in the regulations. Measure similarity between words in Indonesian dictionary with words in the regulation performed to obtain the relationship. Then, words in Indonesian regulations connected with words English regulations using legal interlingual link.

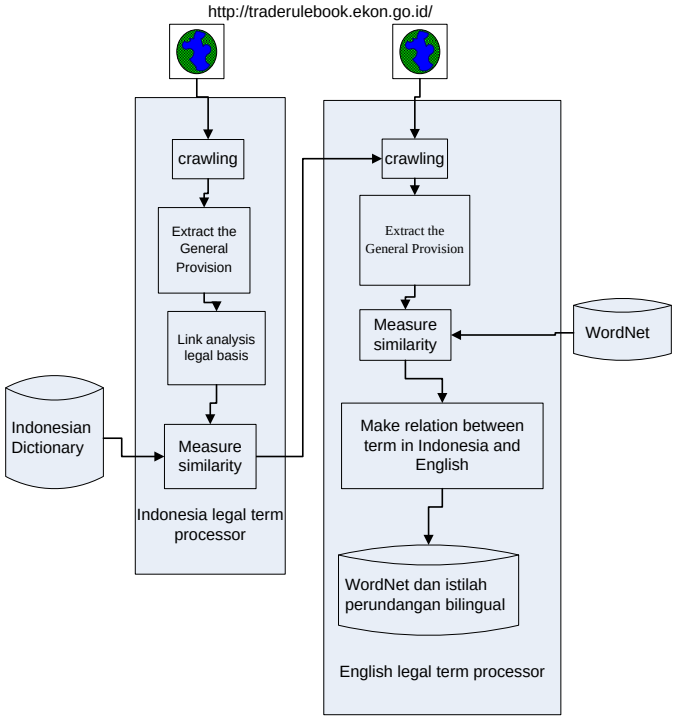

Fig. 1: Word extraction process from legal document

Lexical database that is built up further represented in the ontology using the Web Ontology Language. Information regulation structures, architectural structures such as WordNet lexical database and the links between languages store in RDF/OWL format. OWL is considered to represent lexical database since OWL can represent the hierarchical structure, represent the link

between one component with the other components, and capable of reasoning. With the ability of reasoning, it can be prevented if there are inconsistencies in constructing ontologies.

Then detail steps to extract word in legal document to enter into lexical database is as follows:

1. Extract the General provision section of Rule.
2. If there are words and definitions, then measure semantic relatedness with entry words and its definition in Kamus Besar Bahasa Indonesia (Indonesian Dictionary) using Lesk algorithm. Lesk algorithm measure assigns relatedness by finding and scoring overlaps between the glosses of the two concepts.
3. If there are similarities, then create a link between the word in legal document with word in a lexical database. Else, inserted as a new word in the lexical database.
4. Other regulations examined whether it containes words. Link analysis carried out in the legal basis to examine the relationship between regulation with other regulations. If word already exists in higher regulations, then word in lower regulations refers to higher. If word already exists in regulations at same height, then the word in the more recent regulations refers to a longer word.

After processing the legal documents in Indonesian, further processing legal document in English.

1. Extract the General Provision section of Rule.
2. If there are words and definitions, then measure semantic relatedness with words and its definition in Princeton WordNet. If there are similarities, then create a link between the word in legislation with words in WordNet. If word is not in WordNet, then inserted as a new word in WordNet.
3. Other regulations examined whether it containes word. Analysis carried out in the legal basis to examine the relationship between regulation with other regulations. If word already exists in higher regulations, then word in lower regulations refers to higher. If word already exists in regulations at same height, then the word in the more recent regulations refers to a longer word.
4. Create a link to word in Indonesian regulation.

Table 1. Part of title and general provision law in Indonesian and English

UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 7 TAHUN 2011 TENTANG MATA UANG	LAW OF THE REPUBLIC OF INDONESIA No. 7/2011 CONCERNING CURRENCY
BAB I KETENTUAN UMUM Pasal 1 Dalam Undang-Undang ini yang dimaksud dengan: 1. Mata Uang adalah uang yang dikeluarkan oleh Negara Kesatuan Republik Indonesia yang selanjutnya disebut Rupiah. 2. Uang adalah alat pembayaran yang sah. 3. Bank Indonesia adalah bank sentral Republik Indonesia sebagaimana dimaksud dalam Undang-Undang Dasar Republik Indonesia Tahun 1945.	CHAPTER I GENERAL PROVISION Article 1 In this Law: 1. Currency is money released by the Unitary State of the Republic of Indonesia hereinafter called as Rupiah. 2. Money is legal payment instrument. 3. Bank of Indonesia is central bank of the Republic of Indonesia as intended in the 1945 Constitution of the Republic of Indonesia.

For example, Table 1 show part of title and general provision from law. Table 2 show terms and definitions results from law extraction in Table 1. Table 2 show that Mata uang and Rupiah in Indonesian are synonym , also Currency and Rupiah in English are synonym.

Table 2. Legal terms and definitions

Kata	Definisi	Word	Definition
Mata Uang, Rupiah	uang yang dikeluarkan oleh Negara Kesatuan Republik Indonesia	Currency, Rupiah	money released by the Unitary State of the Republic of Indonesia
Uang	alat pembayaran yang sah.	Money	legal payment instrument
Bank Indonesia	bank sentral Republik Indonesia sebagaimana dimaksud dalam Undang-Undang Dasar Republik Indonesia Tahun 1945	Bank of Indonesia	central bank of the Republic of Indonesia as intended in the 1945 Constitution of the Republic of Indonesia

542 files Indonesian regulations and 405 files English regulations extracted to get words and their definition. There are 363 documents Indonesian regulations that have 3679 words and their definitions. But from 3679, there are distinguishable 2179 words. This is because sometimes word redefine at regulations in lower hierarchy.

Table 3. Indonesian word and their definitions in Indonesia legal document

No	Regulation type	regulation	Word and definition	Average
1	Law	11	268	24,36
2	Government Regulation	13	166	12,77
3	Others	339	3245	9,57

From 405 files English regulations, there are 233 documents that have word and its definition. There are 2203 words and their definitions in Indonesian and English get from 233 documents.

Table 2. English word and their definitions Indonesia legal document

No	Regulation type	regulation	Word and definition	Average
1	Law	11	268	24,36
2	Government Regulation	13	166	12,77
3	Others	209	1769	8,46

From Table 3 and Table 4, it can be seen that the law has more words and definitions. This is because law covers a wider aspect than the regulation lower hierarchy.

4. Architecture Ontology Lexical Database

The main task of this research is develop relation between words and their definition in Indonesia regulations both in Indonesian and English, Kamus Besar Bahasa Indonesia (Indonesian Dictionary) and Princeton WordNet based on the EuroWordNet (EWN) framework [15]. There are challenge in constructing ontology as result of distinguish lexical information and legal information. The challenges are how ontology store lexical information and legal information. Lexical information consist of terms , lexical meanings assigned to them and part

of speech. Legal information consist of hierarchy of regulations, type regulations, terms, and their definitions from legal documents.

WordNet is an initiative of the linguist George Miller and was developed and is being maintained at Princeton University [8]. It encompasses an English-language electronic lexical database inspired by psycho-linguistic and computational theories of human lexical memory. A WordNet serves to support automatic text analysis and AI applications, and to provide an intuitively usable enhanced dictionary.

EuroWordNet is a multilingual lexical database with wordnets for several European languages, which are structured along the same lines as the Princeton WordNet. The most important difference of EuroWordNet with respect to WordNet is its multilinguality [15]. Inter-Lingual-Index made explicit as equivalent relations between the synsets in different languages and WordNet. Each synset in the monolingual wordnets has at least one equivalence relation with a record in this ILI, either directly or indirectly via other related synsets. Language-specific synsets linked to the same ILI-record should thus be equivalent across the languages.

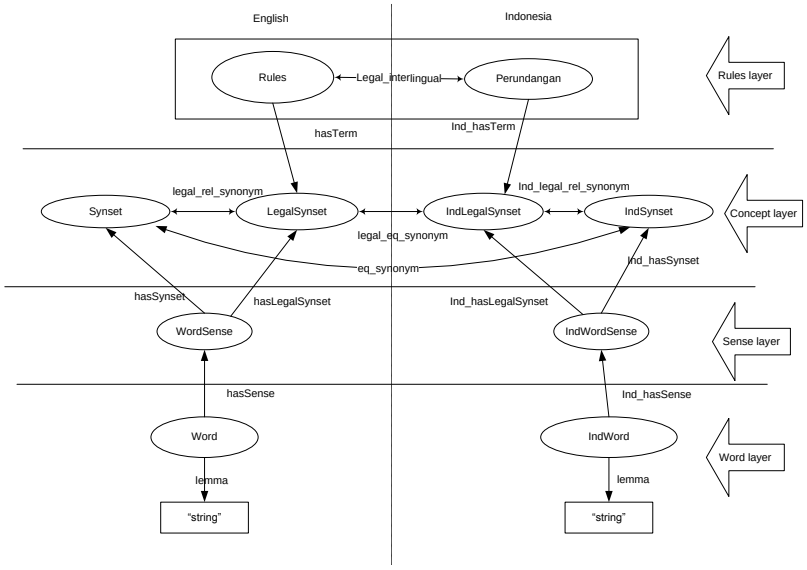

Fig. 2. Schema of OWL representation

Many researchs has been done to convert from fromWord-Net's Prolog format to RDF/OWL which differ in design choices and scope. The main motivations on the development of OWL representation for WordNet lie in two aspects. First, an OWL representation of WordNet provide reasoning procedure. Second, application of WordNet for such tasks on the Semantic Web requires a representation of WordNet in RDF and/or OWL [13][6].

Special structures needed to relate and unify Princeton WordNet-Indonesia lexical database in a multilingual lexical resources. Model representation of Indonesian lexical database follows the model developed by [2] and [11]. Figure 2 presents OWL structure divide into four layers, namely, Word Layer, Sense Layer , Concept Layer and Rules layer.

Rules layer is added to accommodate legal information. Rules layer contain information about hierarchical structure of rules and information about Rules like title, date of adoption and those who authorize. In this layer also create Legal Inter Lingual link that connects between Rule in Indonesian with Rule in English. Structural design using the class hierarchy Protege shown in Figure 3. In Figure 3 may look lexical information and legal information compiled in the class.

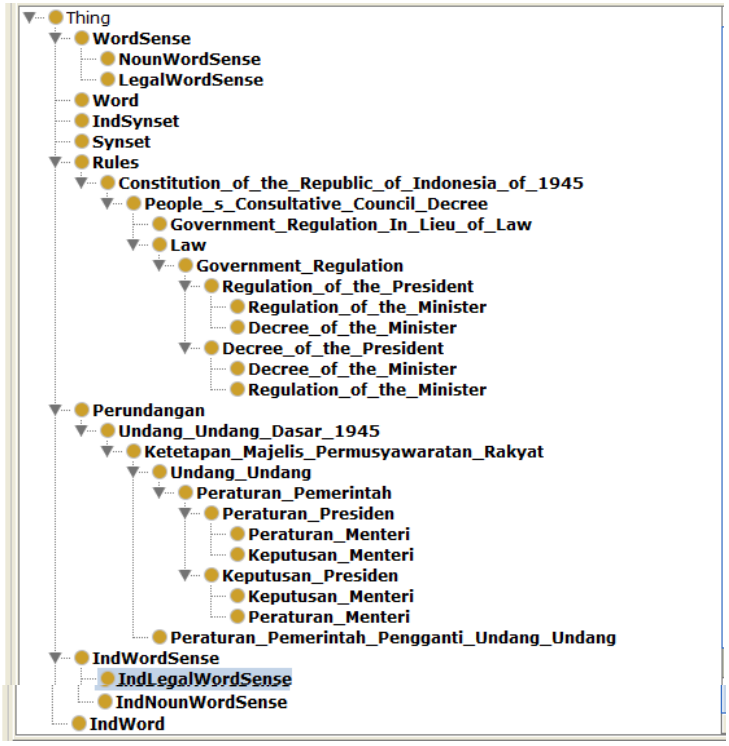

Fig. 3. Hierarchy class in RDF/OWL

Properties defined in Table 3 and 4 adapted from [2][11]. Some properties are added to accommodate the link between language and legal information. Tables 3 and 4 list the datatype properties and object properties defined in this schema respectively. Datatype properties describe an attribute of classes in the form of XML Schema Datatypes (e.g. *legalid*).

The table Object property splits the properties into four categories: properties that
• connect inter lingual (e.g. *legal_eq_synonym*);
• connect the main classes to each other (e.g. *hasTerm*);
• represent relations between IdSynsets (e.g. *hyponymOf*);
• represent relations between Rules (e.g. *legal_basis*).

Table 3. Datatype property in RDF/OWL

Property name	Domain	Range
synsetid	Synset	string
gloss	Synset	string
senseNumber	WordSense	string
lemma	Word	string
Ind_synsetid	IndSynset	string
Ind_gloss	IndSynset	string
legal_ gloss	LegalSynset	string
legal_synsetid	LegalSynset	string
Ind_legal_ gloss	IndLegalSynset	string
Ind_legal_synsetid	IndLegalSynset	string
legalid	Rules	string
legaltitle	Rules	string
Ind_legalid	Perundangan	string
Ind_legaltitle	Perundangan	string
IndsenseNumber	IndWordSense	string

lemma	IndWord	string

Table 4. Object property in RDF/OWL

Property Name	Domain	Range
legal_interlingual	Rules	Perundangan
eq_synonym	NounSynset	IndNounSynset
legal_eq_synonym	LegalSynset	LegalSynsetId
hasTerm	Rules	LegalSynset
Ind_hasTerm	Perundangan	IndLegalSynset
hasLegalSynset	WordSense	LegalSynset
Ind_hasLegalSynset	IndWordSense	IndLegalSynset
Ind_hasSense	IndWord	IndWordSense
legal_rel_synonym	Synset	LegalSynset
Ind_legal_rel_synonym	IndSynset	IndLegalSynset
Ind_hasWord	IndWordSense	IndWord
Ind_hasSynset	IndWordSense	IndSynset
hyperonym	IndNounSynset	IndNounSynset
hyponym	IndNounSynset	IndNounSynset
Ind_legal_basis	Perundangan	Perundangan
Ind_legal_basis_for	Perundangan	Perundangan
legal_basis	Rules	Rules
legal_basis_for	Rules	Rules

Synset used for storing synset lexical information. *LegalSynset* used for storing synset legal information. *legal_rel_synonym* link is used to connected synset lexical information in *Synset* with synset legal information in *LegalSynset*. *hasTerm* link is used to connected *Rules* with synset *LegalSynset*. This link means that legal synset exist in *Rules*. Synset in English and synset in Indonesia that have the same sense made relations *eq_synonym*. *legal_eq_synonym* is used to made relations synset legal information in *LegalSynset* and synset legal information in *IndLegalSynset* that have the same sense. Relation between Rules in English with Perundangan in Indonesian connected use *legal_inter_lingual* link. *legal_basis_for* and *legal_basis* link means that there are legal basis relation between regulation.

5. Conclusion

Extraction of words and definitions in part the general provisions of the regulation becomes the basis for building a dictionary of legal terms. Link analysis is done on the legal basis to see if there are relationship between words and their definition in the regulation with other regulations. This method is able to look for consistency of words and definitions based on the hierarchy of regulation.

Extraction results showed that the higher regulation has more words and definitions. This is because the higher regulation include wider aspects than the regulations lower hierarchy. Extraction on Indonesian and English regulation documents produce legal bilingual dictionary Indonesian and English.

Ontology architecture presented in this paper is represent knowledge about the legal information in Indonesia legal domain. The main components of lexical information and the hierarchy of rules transformed as classes in OWL. Relations between synset, lexical, legislation transformed as OWL properties.

Lexical representation is storing information in Indonesia Dictionary and Princeton WordNet as well as words and their definitions in legal document. Also, representations are store information words and definitions in the regulation which are associated with the hierarchy of regulation.

REFERENCES

[1.] Arora, Pooja,(2012), *Semantic Searching and Ranking of Documents using Hybrid Learning System and WordNet*, International Journal of Advanced Computer Science and Applications, Vol. 3, No. 6

[2.] Assem, M. Van, Gangemi, A. and Schreiber, G.,(2001), *Conversion of WordNet to a standard RDF / OWL representation*, In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy, pp. 237–242.

[3.] Banerjee, S., and Pedersen, T. (2003), *Extended gloss overlaps as a measure of semantic relatedness*, In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, 805–810.

[4.] Breuker, J., Valente, A. and Winkels, R., (2004), *Legal ontologies in knowledge engineering and information management*, Artificial Intelligence and Law, 12(4), pp.241–277.

[5.] Despres, S. and Szulman, S., (2004), *Construction of a Legal Ontology from a European Community Legislative Text, Legal Knowledge and Information Systems*, Jurix 2004: The Seventeenth Annual Conference, pp.79–88.

[6.] Dean, M. A., Schreiber, Th. , Bechofer, S. , F. van Harmelen, J. Hendler, I. Horrocks, D. MacGuinness, P. Patel-Schneider, and L. A. Stein, (2004), *OWL Web Ontology Language Reference*, W3C Recommendation, World Wide Web Consortium, 10 February, Latest version: http://www.w3.org/TR/owl-ref/.

[7.] Dini, L., Peters, W., (2005), *Cross-lingual legal information retrieval using a WordNet architecture*, Proceedings of the 10th international conference on Artificial intelligence and law - ICAIL '05, p.163.

[8.] Fellbaum, C., (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, London, England, 71(3), p.423.

[9.] Gangemi, A.,Sagri, M.-T.,Tiscornia, D.,(2003),*Metadata for content description in legal information*,14th International Workshop on Database and Expert Systems Applications Proceedings.

[10.] Gruber, T.R. ,(1993), *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, Technical Report KSL-93-4, Knowledge Systems Laboratory, Stanford University, Stanford, United States.

[11.] Huang, X. dan Zhou, C., (2007), *An OWL-based WordNet lexical ontology*, Journal of Zhejiang University SCIENCE A, 8(6), pp.864–870.

[12.] Kharkevich, U., (2010), *Concept Search : Semantics Enabled Information Retrieval*. PhD Dissertation, International Doctorate School in Information and Communication Technologies, University of Trento, (March).

[13.] Manola, F., & Miller, E. ,(2004), *RDF primer: W3C recommendation*.

[14.] Peters, W., Sagri, M.T. dan Tiscornia, D., (2007), *The structuring of legal knowledge in LOIS*, Artificial Intelligence and Law, 15(2), pp.117–135.

[15.] Vossen, P., (1997), *EuroWordNet: a multilingual database for information retrieval*, In Proceedings of the DELOS workshop on Cross-language Information Retrieval, pp.5–7.

[16.] http://portal.mahkamahkonstitusi.go.id/eLaw/about_us.php

[17.] Law Of The Republic Of Indonesia No. 12 Of 2011 Concerning Making Rules, http://traderulebook.ekon.go.id/ 4778_UU_12_2011_e.html

[18.] Oxford Advanced Learner's Dictionary, 8th Edition.