An Automatic Approach for Bilingual Tuberculosis Ontology Based on Ontology Design Patterns (ODPs)

Bambang Harjito*, Denis Eka Cahyani, Afrizal Doewes
[1,2,3]Sebelas Maret University, Department of Informatics Faculty of Mathematics & Natural Sciences,
Surakarta, 57126, Indonesia
Corresponding author, e-mail: bambang_harhito@staff.uns.ac.id, denis.eka@staff.uns.ac.id,
afrizal.doewes@staff.uns.ac.id

Abstract

Ontology is a representation term used to describe and represent a domain of knowledge. Manually ontology development is currently considered complex, requiring a lot of time and effort. This research was proposed to develop methods to build automatic domain ontology bilingual in Indonesian and English by using corpus and ontology design patterns (ODPs) in tuberculosis disease. In this study, the methods used were to combine ontology learning from text and ontology design patterns to decrease the role of expert knowledge. The methods in this research consist of six stages are term and relation extraction, matching with Tuberculosis glossary, matching with ODPs, score computation similarity term and relations with ODPs, ontology building and ontology evaluation. The results of ontology construction were 362 terms and 44 relations with 260 terms were added. The calculation accuracy of ontology construction was 71%. Ontology construction had higher complexity and shorter time as well as decreases the role of the expert knowledge which proof that the automatic ontology evaluation is better than manual ontology construction.

Keywords: *automatic, ontology building, ontology design patterns, tuberculosis*

1. Introduction

Tuberculosis is a public health problem in the world. Tuberculosis (TB) is an infectious bacterial disease caused by the microorganism Mycobacterium tuberculosis that affected the human lungs but can also on the organ or other tissue such as skin, eye, lymph nodes, bone, a lining of the brain and other organs [1, 2].

World Health Organization (WHO) estimated that 8.7 million new cases and 1.4 million died of tuberculosis cases annually. Approximately 75% of patients Tuberculosis were in the most productive age (15-50 years). Other than economic disadvantages caused by the lost of annual income over the patient, tuberculosis had another negative impact such as social stigma and even ostracized by the community. Indonesia was ranked fourth of the most amount of tuberculosis cases in the world after India, China, and South Africa [3].

One way to prevent the growth of patients suffering from Tuberculosis disease is to improve the quality of capable health workers to handle the tuberculosis disease situation. The qualities of health workers can be improved by increasing their knowledge of tuberculosis cases in the society. Increasing knowledge of health workers against disease will impact the health services to be better for society.

By the development of technology, sources of knowledge about tuberculosis disease can be obtained easily from textbooks, scientific journals, websites etc. Currently, there are several websites which publish a collection of scientific journals on health, including Tuberculosis in Indonesia, e.g. Health Science Journal of Indonesia and Makara Journal of Health Research. These websites have hundreds of scientific journals related to health, including the Tuberculosis disease that can be utilized to increase knowledge of health workers in managing tuberculosis disease situation [4].

Scientific journals are a source of knowledge that is vital to develop research and technology regarding the disease in Indonesia, including tuberculosis. Text in the scientific journal can be used to build ontology in health, particularly tuberculosis disease. Ontology is a representation term used to describe and represent a domain of knowledge [5]. Ontology as a

Received October 5, 2017; Revised December 10, 2017; Accepted December 29, 2017

knowledge representation method can effectively represent the concepts of structure and the relations between concepts [6]. Ontology languages express a rich semantic and provide best reasoning capabilities [7]. Building ontology can be a representation of knowledge over information about the tuberculosis disease. One part of the scientific journal is abstract in form Indonesian and English language which was used as corpus resource in building ontology.

Besides using corpus, the development of ontology can also use ontology design patterns (ODPs). Ontology design patterns constituted derivative of the design patterns used in software engineering. Ontology design patterns were defined as a pattern to identify the ontology structure design. Design patterns set aside the dependencies between terms so that if there was a change in the terms, it would not affect the other terms [7]. The use of Ontology Design Patterns (ODPs) has been shown to have beneficial effects on the quality of developed ontologies, and promises increased interoperability of those same ontologies [8]. Ontology design patterns (ODPs) are a proposed solution to facilitate ontology development, and to help users avoid some of the most frequent modeling mistakes [9]. The ontology design patterns also offer advantages enabling a more modular, well-founded and richer representation of the knowledge. This representation will produce a more efficient knowledge management in the long term [10].

Based on this background, it is important to do research related to the development of ontology domain bilingual corpus of scientific journals and ontology design patterns to represent knowledge [11]. The manual construction of ontology that had been done before was too complex, requiring a lot of time and effort [12]. Therefore, an automatic process is needed to facilitate the development of ontology. The existing approaches of automatic processes is ontology design patterns (ODPs) [13].

The main contribution of this paper is present ontology development with automatic approach using ontology design patterns (ODPs) for tuberculosis domain. This paper improves the results of research previously. Drame, et al., proposed to develop a semi-automatic ontology building in Alzheimer domain using corpus and bilingual UMLs Meta thesaurus [14]. Validation of the ontology used by the expert was to ensure the knowledge in ontology. However, it took about one month to validate the ontology. Therefore, this paper was developed using the corpus and ODPs, so that validation can be done without expert. Dahab, et al., [15] build automatic construction ontology from natural language text using semantic pattern approach. Then, Navigli and Velardi [16] developed a methodology for automatic ontology enrichment and document annotation. Natural language definitions from available glossaries were processed and regular expressions are applied to build the ontology. This paper was different from their studies [14, 15] because it used bilingual corpus and ontology design patterns (ODPs) approach for building ontology automatically. Mortensen, et al., proposed applications of ontology design patterns (ODPs) in Biomedical Ontologies [9] and Cahyani, et al., [17] also purposed development ontology using ontology design patterns (ODPs) in Alzheimer domain, but in this paper we show the utilization of ODPs to bulid ontology automatically in tuberculosis disease.

2. Resources Used

The resource was divided into data and tools to process terminological resources.

2.1. Data

2.1.1. Corpus

The corpus used in this research was the abstract (English and Indonesia language) in group health scientific journals in websites such as Health Science Journals, Health Science Journal of Indonesia, Makara Journal of Health Research. Corpus abstract of a scientific paper from the journal had enriched knowledge about Tuberculosis. Currently, there were 55 papers published in this scientific journal and reviewed by expert research domain.

2.1.2. Tuberculosis Glossary

This research used a glossary term to filter the results of a Tuberculosis extraction from the corpus. The filtering term extraction results were needed to get a term linked to the Tuberculosis disease. Tuberculosis's glossary obtained at the website address (http://www.tbindonesia.or.id/). The total of terms which were related to Tuberculosis disease in this glossary was 840 terms.

2.1.3. Ontology Design Patterns (ODPs)

Ontology Design Patterns (ODPs) could be accessed at http://www.gong.manchester.ac.uk/odp/html/index.html. This website also contained a catalog of ODPs. In this catalog, there were three types of ODPs; (i) Domain Modeling ODPs, (ii) Good Practice ODPs, and (iii) Extension ODPs. The total number of ontology design patterns in the catalog was 17 ODPs. ODPs Domain Modeling aimed to get the best model for specific domain ontology, e.g. Interactor_Role_Interaction and Sequence. Good Practice design pattern ontology aimed to be better and stronger to maintain ontology models, e.g. Normalization and Upper-Level Ontology. On the other hand, Extension design pattern aimed to overcome the limitations of existing ontology models to expand or increase coverage of the ontology, e.g. Nary_Data Type Relationship and Exception.

2.2. Tools

2.2.1. Text2Onto

Text2Onto is a framework of learning ontology which developed to support ontology construction from textual documents. Text2Onto has been used by Cimiano and Volker [12]. The research used Text2Onto as a framework for ontology learning from textual resources based on Probabilistic Ontology Model (POM). There were three processes in Text2Onto: preprocessing, Executing of Algorithms and Combining results. During preprocessing, Text2Onto called GATE application to tokenize document and tag Part of Speech sentences to create indexes for the document, and the result of this process was obtained as an annotation document. Executing of Algorithms was the process of Text2Onto executed the applied algorithms to extract terms and relations. One of the applied algorithms was TFIDF Concept Extraction. The last process was combining results; this process combined the result of extracted terms and relations derived from processed documents. Text2Onto was available at http://code.google.com/p/text2onto/downloads/list.

2.2.2. SimMetrics

SimMetrics is an open-source library available in Java, which contains more than 20 similarity distance algorithm, e.g. Jaro-Winkler, Levenstein distance, and Monge-Elkan distance. SimMetrics used for string correspond to identify the position of string or set of strings within a text. String correspond algorithms compared two different strings and found the similarity score between two text comparisons. SimMetrics has been used by Chapman, et al., [18]. This research used SimMetrics to calculate the similarity between texts, where the information in this text had been integrated into large repositories (e.g. the Web). SimMetrics was available at https://github.com/Simmetrics/simmetrics.

2.2.3. Ontology Generation

Ontology generation is a plug-in protégé to build ontology with generating terms of natural language text. Ontology generation was developed by Watcher and Schroeder, 2010 [19]. This tool supported the creation and extension of OBO ontology by semi-automatically generating terms, definitions, and parent-child relations from the text in PubMed; the web, and PDF repositories. This tool generated term by identifying significant noun phrases in text statistically and for the definitions and parent-child relations, it employed pattern-based web searches. Ontology generation was available at http://protegewiki.stanford.edu/wiki/Ontology_Generation_Plugin_(DOG4DAG). Ontology generation can be applied to the protégé-OWL version 4.3.

3. Research Method

The methods in this research consist of six stages: (a) Term and relation extraction (b) Matching with Tuberculosis glossary (c) Matching with ontology design patterns (ODPs) (d) Score computation similarity term and relations with ODPs (e) Ontology building (f) Ontology evaluation. The process of each stage in the method in this research was in Figure 1.

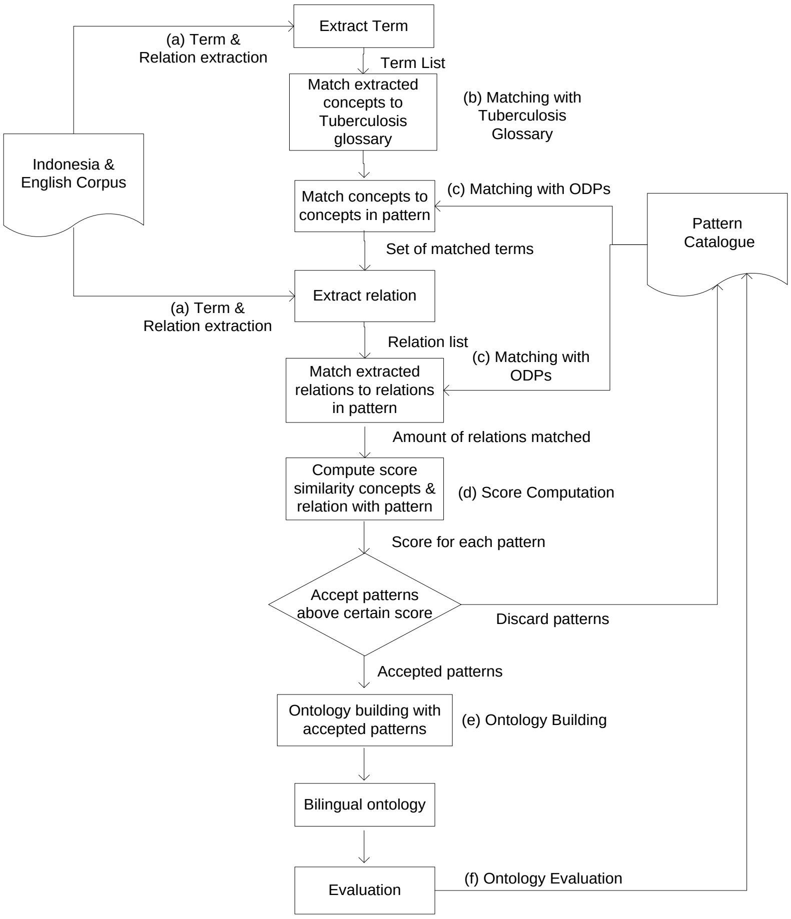
Extract Term
(a) Term & Relation extraction
Term List
Match extracted concepts to Tuberculosis glossary
(b) Matching with Tuberculosis Glossary
Indonesia & English Corpus
Match concepts to concepts in pattern
(c) Matching with ODPs
Pattern Catalogue
Set of matched terms
Extract relation
(a) Term & Relation extraction
Relation list
(c) Matching with ODPs
Match extracted relations to relations in pattern
Amount of relations matched
Compute score similarity concepts & relation with pattern
(d) Score Computation
Score for each pattern
Accept patterns above certain score
Discard patterns
Accepted patterns
Ontology building with accepted patterns
(e) Ontology Building
Bilingual ontology
Evaluation
(f) Ontology Evaluation

Figure 1. Overview of ontology building method

The main idea of this research was to extract terms and associations and then Correspond it on design patterns. Then build the ontology and enrich them with parallel corpus. The stages of the method in this paper were explained as follows.

a) Term & relation extraction

Corpus of Health Science Journals in English and Indonesia were extracted to retrieve a number of the terms. Next, the extraction of relationships linked between terms to retrieve a number of relations. This corpus extraction used was Text2Onto.

b) Matching with tuberculosis glossary

After obtaining a number of terms and relations, the next stage was matching with the glossary of Tuberculosis. At this stage, the matching of Tuberculosis glossary aimed to filter terms in order to derive from the extracted word list and in the glossary.

c) Matching with ontology design patterns

The extracted terms and relations compared with terms and relations contained in Catalog ODPs that consist 17 design patterns. The matching result was calculated to get the score of similarity by using SimMetrics tools that used Euclidean Distance algorithm. Then, two scores obtained from correspond concepts and relations were weighted together to form a “total-matching-score” for each pattern. Then a decision was made according to some threshold value, the patterns were kept and included in the ontology result, which would be discarded. Finally, an ontology was built from the accepted patterns which have the highest score of similarity.

d) Score Computation

At this stage, the similarity calculation computed between the extracted term and relation of the concepts and the relationships that exist in the design pattern. The tool used was SimMetrics which consisted of various algorithms, e.g. Euclidean Distance similarity distance, Levenshtein, and so on. So, average values were calculated from all the existing algorithms, to obtain value or score for string matching. The result was the value or similarity score for each design pattern. Afterward, a design pattern that has the highest similarity score was implemented to build ontology. More attention was given for relation between the concepts because it was capable of making more structural ontology.

e) Ontology Building

Ontology building was the stage to build ontology of terms and relations that correspond to ontology design patterns. The ontology constructed implement design pattern that has the highest similarity values on the ontology of Tuberculosis that was built. This stage used OWL ontology generation to build ontology from terms and relationships that exist. The first step to use ontology generation was search definition of the term entered. The search was connecting with PubMed in the protégé. Then, the automatic map of terms and relation existed as to build a new ontology.

f) Ontology Evaluation

Ontology evaluation was viewed in terms of complexity, time and effort required to build this ontology. Moreover, ontology evaluation also calculated the accuracy of the terms and relation that used to build the ontology. Accuracy is calculated by the following formula:

$$accuracy = \frac{x}{y} \tag{1}$$

where; x=matching results of term/relation
y=total all of match term/relation

x was the matching results of term or relation suitable terms and relations extracted from the corpus and in design patterns that have the highest score similarity.

Meanwhile, y was the total all term/relation extracted from the corpus and had been filtered by Tuberculosis's Glossary. Those terms and relations were corresponded with the terms and relation on ODPs.

4. Result and Analysis

This section is about the results of the steps for building a fully automatic ontology construction.

a) Term & relation extraction

The results at this stage were a collection of terms and relationships from corpus extracted by using Text2Onto. The corpus used in this study was 55 papers. The results obtained were 1310 terms and 44 relations between terms. The number of the terms resulting from the extraction of corpus turns out quite a lot, so it was needed to be filtered to get the appropriate terms that related to the Tuberculosis disease.

b) Matching with Tuberculosis Glossaries

The result of terms and relations extraction of this stage was filtering by matching Tuberculosis's glossary contained 860 terms related to Tuberculosis which acquired 260 matching terms. This was different from the terms extracted from a corpus using extraction with Text2Onto because the extraction term related many health terms in general, not specifically related to Tuberculosis disease. In addition, the number of terms in the Tuberculosis's was less than terms of general health glossary so the scope of term filtering would be limited.

c) Matching with ontology design patterns

Terms and relations that had been filtered would be corresponded with a list of terms and relations that exist in the ontology design patterns. In the catalog, there were several kinds of ontology design patterns (ODPs). The corresponded results were calculated for the similarity values between terms and filtered results with a term relation and relation that exist in the ontology design patterns (ODPs).

d) Score computation

The result of similarity matching between term and relation with each ontology design patterns are shown in Table 1.

The highest value of similarity found in ontology design patterns closure was equal to 81%. Closure ontology design pattern was a design pattern that limits the relationships among concepts which allowed it to happen by clarifying the relation [20]. The limitations in this relation were to express a concept has had a particular relation and only those relations, e.g. a carnivorous is a meat eating animals, with closure design pattern was revealed that carnivores do not eat other foods besides meat.

Table 1. Result of the Similarity Calculation ODPs

No	ODPs Type	Name	Similarity score
1	Domain_Modelling_ODP	Adapted_SEP	80%
2		CompositePropertyChain	80%
3		Interactor_Role_Interaction	79%
4		List	76%
5		Sequence	78%
6	Extension_ODP	Exception	80%
7		Nary_DataType_Relationship	80%
8		Nary_Relationship	79%
9	Good_practice	Closure	81%
10		DefinedClass_Description	80%
11		Entity_Feature_Value	76%
12		Entity_Property_Quality	80%
13		Entity_Quality	80%
14		Normalisation	80%
15		Selector	79%
16		Upper_Level_Ontology	78%
17		Value_Partition	80%

e) Ontology Building

The ontology built in this research consisted of several components; there were 362 terms and 44 relations. Terms and relations used to build the ontology was OWL ontology generation. There were 260 new terms added in that ontology. Figure 2 represented the results of the ontology that has been built in the protégé editor tool.

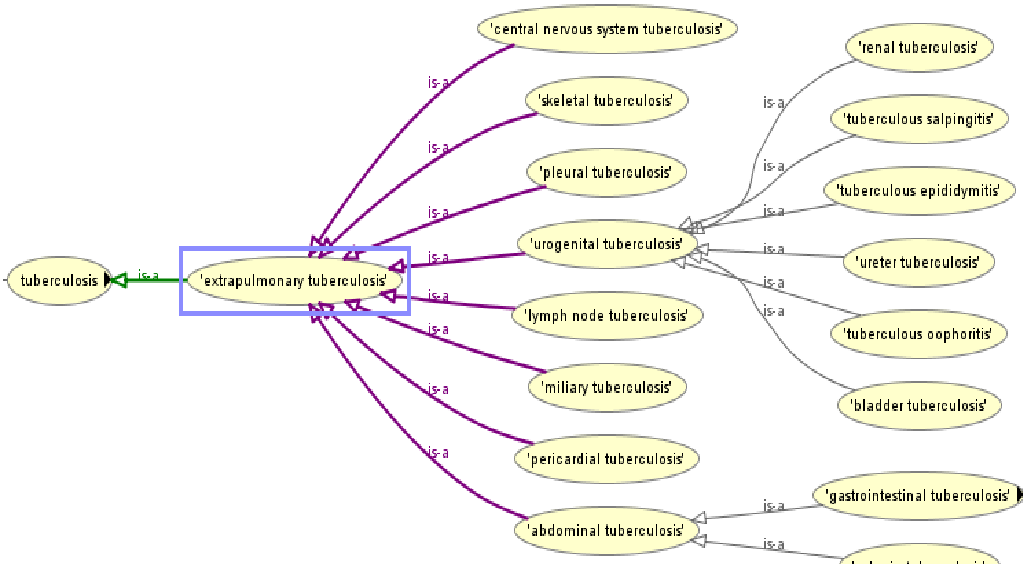
'central nervous system tuberculosis'
'renal tuberculosis'
'skeletal tuberculosis'
'tuberculous salpingitis'
'pleural tuberculosis'
'tuberculous epididymitis'
'urogenital tuberculosis'
tuberculosis
is-a
'extrapulmonary tuberculosis'
'ureter tuberculosis'
'lymph node tuberculosis'
'tuberculous oophoritis'
'miliary tuberculosis'
'bladder tuberculosis'
'pericardial tuberculosis'
'gastrointestinal tuberculosis'
'abdominal tuberculosis'
'splenic tuberculosis'

Figure 2. Visualization a part of the ontology in protégé

f) Ontology Evaluation

The result of accuracy value of fully automatic ontology construction was 71%. It was obtained from the calculation of matching number of 260 terms or relations and the total term or relation in the ontology built 362 terms or relations. This indicated that fully automatic ontology construction method used in the study was quite excellent to be able to build the ontology. The results of accuracy in this study were similar to those of previous studies [17] in the domain of Alzheimer's disease that resulted in 71% accuracy. The accuracy value can be as a supporting material to the evaluation of this research.This indicates the method of ontology construction that use Ontology Design Patterns (ODPs) in this research can be applied to various domains and get good accuracy value on the result.

The evaluation of ontology that has been built can be seen in terms of the complexity, time and effort required. The results of automatic ontology construction were able to shorten the time when compared to the construction of ontology manually or semi-automatic that required validation of at least one-month length of an expert [14]. In this study we just need several days to validation the ontology with Ontology Design Patterns (ODPs). This indicated the method in this paper can save time to building the ontology. In previous studies [14] it takes two teams in the field of Alzheimer's expert to validate the built of ontology. While in this study does not require expert to validate ontology so we can save effort to build ontology automatically. So in this paper, we have the advantage of time and effort required aspect to build ontology construction.

5. Conclusion and Future Work

This research succeeded to make fully automatic bilingual domain ontology using the Ontology Design Patterns (ODPs) and corpus. The result of ontology development included 361 terms and 44 relations with the addition of 260 terms. The calculation accuracy of ontology construction was 71%.Fully automatic construction could speed up and decrease the human's role as the expert to evaluate ontology rather than building ontology manually. The result of the evaluation was fully automatic ontology constructions that shorten development time compared to manual ontology or semi-automatic which required expert validation.

For future work, it is suggested to add more terms and relation in Tuberculosis’s glossary in order to have well filtered terms results of corpus extraction. In addition, type of data ontology design patterns (ODPs) can be improved to get the highest similarity value for selected design patterns that implemented to build the ontology. Moreover, ontology enriches the number of terms in order to be implemented in ontology building. Ontology enrichment using parallel corpus of the website in English and Indonesia can obtain terms and synonymous terms in other languages.

References

[1] Kemenkes. National Guidelines For The Control Of Tuberculosis. Directorate General of disease controls and environmental health. The Ministry Of Health Of Indonesia. Jakarta. 2014.

[2] World Health Organization. Definition and Reporting Framework for Tuberculosis – 2013 revision. Geneva: WHO Press. 2013.

[3] World Health Organization. Global Tuberculosis Report 2016. Geneva: WHO Press. 2016.

[4] Gizaw GD, Alemu ZA, Kibret KT. Assessment of knowledge and practice of health workers towards tuberculosis infection control and associated factors in public health facilities of Addis Ababa, Ethiopia: A cross-sectional study. The official journal of the Belgian Public Health Association. 2015; 73(15).

[5] Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition*. 1993; 5: 199-220.

[6] Eutamene A, Kholladi MK, Belhadef H. Ontologies and bigram-based Approach for Isolated Non-word Errors Correction in OCR System. *IJECE International Journal of Electrical and Computer Engineering*. 2015; 5(6): 458-1467.

[7] Gan J, Xie G, Yan Y, Liu W. Heterogeneous Information Knowledge Construction Based on Ontology. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2016; 14(4): 1617-1628.

[8] Louis Jean L. Prototype System For Automatic Ontology Construction. Thesis Magister Information Technology. Sweden: The Royal Institute Of Technology; 2007.

[9] Hammar K. *Ontology Design Patterns in WebProtege*. CEUR Workshop Proceedings. 2015; 1486.

[10] Mortensen JM, et al. *Applications of Ontology Design Patterns in Biomedical Ontologies*. AMIA Annual Symposium Proceedings. 2012: 643–652.

[11] Aranguren ME, Antezana E, Kuiper M, Stevens R. *Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology*. BMC Bioinformatics Proceedings. 2008.

[12] Cimiano P, Völker J. *Text2Onto a framework for ontology learning and data-driven change discovery.* Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems NLDB. Alicante, Spain, Springer. 2005; 3513: 227-238.

[13] Blomqvist E. *Fully Automatic Construction of Enterprise Ontologies Using Design Patterns: Initial Method and First Experiences.* In Proceedings of OTM 2005 Conferences, Ontologies, DataBases, and Applications of Semantics (ODBASE). Agia Napa, Cyprus. 2005.

[14] Dramé K, et al. Reuse of terminal-ontological resources and text corpora for building a multilingual domain ontology. *An application to Alzheimer's disease: J Biomed Inform*. 2014.

[15] Dahab MY, Hassan H, Rafea A. TextOntoEx: Automatic ontology construction from natural English text. *Expert System Applications.* 2008; 34: 1474-1480.

[16] Navigli R, Velardi P. From Glossaries to Ontologies : Extracting Semantic Structure from Textual Definitions. *Ontology Learning Population Bridging Gap between Text Knowledge.* 2008; 71-87.

[17] Cahyani DE, Wasito I. Automatic Ontology Construction Using Text Corpora and Ontology Design Patterns (ODPs) in Alzheimer's Disease. *Jurnal Imu Komputer dan Informasi (Journal of Computer Science and Information)*. 2017; 10(2): 59-66.

[18] Chapman S, B Norton, F Ciravegna. *Armadillo: Integrating knowledge for the semantic web.* Proc. DagstuhlSemin. Mach. Learn. Semant. Web. 2005: 2-4.

[19] Wächter T, M Schroeder. Semi-automated ontology generation within OBO-Edit. *Bioinformatics.* 2010; 26: 88-96.

[20] ODP public catalog. Closure. http://www.gong.manchester.ac.uk/odp/html/Closure.html. Access on Monday, 26 May 2014.