An Ontology-Based Representation of Financial Criminology Domain Using Text Analytics Processing

Zulazeze Sahri†, Shuhaida Mohammed Shuhidan[2] and Zuraidah Mohd Sanusi[3],

Universiti Teknologi MARA[1], Accounting Research Institute, Universiti Teknologi MARA Selangor[23]

Summary
Financial Criminology is an emerging field of today's research area in combating and detecting any misuses or fraudulent in financial management. The study of financial criminology by academia and industry is dramatically increased. Therefore, domain understanding and knowledge capturing in financial criminology is needed to help the future researcher to get a better understanding of the domain. This paper proposes an ontology-based representation of financial criminology to capture the commonly talked terms and topics in financial criminology researches and studies. Ontology is one of the knowledge representation techniques that allow us to understand a particular domain (Financial Criminology) in the form of classes (parent) and attributes (children) hierarchy. Twenty Five (25) journals and research papers in financial criminology has been selected for this research in order to extract the commonly talked terms and topics of financial criminology studies. Text Analytics processor tool by RapidMiner has been used to extract and identify the terms, then the tool will analyze the number of frequency of the terms used in each of the research paper. Finally, the identified terms is converted to an ontology representation language (OWL) by using Protégé. The research found that there are nine (9) classes (Topics) that are commonly researched on the field of Financial Criminology. The ontology representation view is validated by three (3) Financial Criminology experts and Auditors. The text analytics result and ontology view of the domain is discussed in the research findings section.

Key words:
Text Analysis, Taxonomy, Domain Understanding, Knowledge Management, Knowledge Representation.

1. Introduction

Financial crime is often defined as crime against property, involving the unlawful conversion of property belonging to another to one's own personal use and benefit. Financial crime is profit-driven crime to gain access to and control over property that belonged to someone else (Gottschalk, 2010). According to Fridson (2002) Financial crimes often involve fraud. They use the terms financial crime, white-collar crime and fraud interchangeably. According to Gottschalk (2010), financial crime can be carried by check and credit card fraud, mortgage fraud, medical fraud, corporate fraud, bank account fraud, payment (point of sale) fraud, currency fraud, and health care fraud, and they involve act such as insider trading, tax violations, kickbacks, embezzlement, identity theft, cyber-attack, money laundering and social engineering. In addition, (INTERPOL, 2017) are focusing on financial crimes in payment cards, money laundering, counterfeit currency and security documents and social engineering fraud. The organization argues the uncontrolled financial crimes will affect individuals, companies, organizations and even nations, and have negative impact on the entire economic and social system through the loss of money incurred.

The important of financial crimes investigation and study has emerge in recent years in commercial and public sectors, including government, enforcement body, law and regulation, private financial sector and also academics. There are many researches in financial crime or white-collar crimes have been conducted these commercial and public sectors. The research fields are various and the findings are highly valuable in combating the frauds. As the knowledge on financial crimes and research fields become emerges, financial crime comes in all kinds of shapes and colours. There are no obvious categorizations or classifications of financial crime has emerged in the literature so far Gottschalk (2010). Therefore, several researchers have study the categorization and classification of financial crime such as (Ngai, Hu, Wong, Chen, & Sun, 2011) and Gottschalk (2010). The studies were focusing on classification into categories of financial crime since the categories will enable a structure within which policing and law enforcement can work. However, research by Gottschalk (2010) doesn't discuss on the methodology used in classification process. In Information Technology, those classification and categorisation of financial crime are presented in the form of semi-structured data and described in natural language. The semi-structured data also cannot be best stored and manipulated in the form of machine or computer readable since it is in the form of taxonomy. Hence, the research interest in how to structure and formulate the financial crime data and natural language (research writing and text) into formal language (machine readable) so that computer can understand.

This paper aims to understand the overall research topics and terms that are commonly used in the financial crime domains. The significant of this research is to give a

Manuscript received February 5, 2018
Manuscript revised February 20, 2018

general overview of financial crime research and can help to capture and structure the domain knowledge into a machine or computer readable format that can help in developing a financial criminal related information management system such as fraud detection and prevention system in near future. In knowledge management, the process of understanding and representing particular domain knowledge is called Knowledge Representation. The knowledge representation scheme can be used to encode and store knowledge in a database or "Knowledge Base" (Rosenberg, 1986). There are several knowledge representation techniques that this research can adapt to structure and model the knowledge. In this research, we are using Ontology as the knowledge representation technique due to its ability to identify class or concept, property and relationship within a domain of discourse. The ontology development involved extracting and capturing the commonly used 'terms' and 'topics' in the field of Financial Crime research findings. A number of Twenty Five (25) research papers, journal and whitepaper has been selected in the Text Mining process in order to extract words (terms) occurrence in all of the research documents. Finally, the text analysis result will be used to model the Financial Crime Ontology using Noy & McGuinness (2001) methods and implemented into Web Ontology Languages (OWL) using Protégé Semantic Editor (OWL Web References). The detail of each technique along with the relevant tools will be explained in the next section.

2. Literature Review

2.1 An Overview of Knowledge Representation

Knowledge representation is an area that explain how a particular domain knowledge can be best represented symbolically and be able to be manipulated in automated way by reasoning program (Computer Program or Information System) (Brachman & Levesque, 2004). In knowledge management, knowledge representation is required to convert tacit knowledge (experience/self-knowledge) to explicit knowledge (documented knowledge), and to represent the explicit knowledge in the form of suitable knowledge representation before it can be modelled and applied in knowledge sharing system such as database, repository or library (Obomsawin, (2001); Grenier, (1998); Mondo et al. (2007)). There are many knowledge representation techniques in AI, the most popular and commonly used are Logic, Production Rules, Semantic Nets and Frame. The most popular and commonly used knowledge representations are Logic which related to the truth of statements about the world (Chakraborty, 2010). It is a formal system in which the formulas or sentences have TRUE or FALSE values only. Secondly is Production Rules knowledge representation technique that relies on IF-THEN rules which be able to provide the flexibility of combining declarative and procedural representation for using them in a unified form. It is popular as a knowledge representation mechanism used in design of many "Rule-Based System" or "Production Systems" (Chakraborty, 2010). Semantic Nets represented knowledge as concept nodes related by directional relationship links, representing the world as a directed graph (Collins & Quillian, 1969). Semantic network commonly used to represent the inheritable knowledge by organized into classes and classes must be arranged in a generalization hierarchy (Poonam et al. 2010). Frame is collection of attributes or slots and their associated values which describe the real world entity. The frame used to represent class (set) and instances (class' elements). It components consists of frame name, attributes (slots) and values. Finally, Yuanyuan, Rujing, Xue, & Yimin (2010) suggest that the Ontology can be used to develop the knowledge based by constructing the ontology model through the relevant concepts and their relationship. Therefore, the formalized ontology knowledge representation makes knowledge sharing and reuse possible through the knowledge based system.

2.2 Knowledge Representation using Ontology

Ontology is a knowledge representation technique that describes a particular domain by defining the concepts of this domain and the relations between them. Noy & McGuinness (2001) defined an ontology as a formal explicit description of concept in a domain of discourse (class or concept), properties that describe the characteristic of the concept (slots or roles), and restriction on slots (facet or role restriction). Generally, ontology as a graph/network structure consisting of;

A set of concepts – Object/Entities (vertices edges in a graph)
A set of relationship connecting concepts (directed edges in a graph)
One set of instances assigned to a particular concepts (data records assigned to concepts or relation)

Inheritance is one the main relations that ontology supports, allowing an IS-A relationship between instances (Nour, 2003). In Information Technology, a complete ontology with a set of individual instances of classes is able to form a knowledge base. For example, based on relations and attributes defined in the ontology description, user can query for a pieces of information related to a particular domain. Then, a query will perform a logical match between relations and attributes specified in the query and those defined in the ontology. The inference engine will perform a logical search using instance

attributes while applying rules wherever necessary in order to evaluate the query (Nour, 2003). In contra with other knowledge representation techniques discussed in previous section, Ontology is the most suitable technique to understand and describe a particular domain such as Financial Criminology because of its ability to identify class or concept, property and relationship within a domain of discourse and finally will allow us to develop a financial criminology related computer program such as fraud detection and prevention system in the future.

2.3 Text Analytics (Tools and Techniques)

The advancement of software development particularly in data science and analytics field brings up the development of text mining and analysis as part of the core processes in big data analysis. In this research, we look at the advance features in data analysis tools and technique that can help in our research work. In relations to ontology development, we found that text mining and analysis can help in the process of extracting and identifying the concepts (terms) used in the research documents with high accuracy and relevancy.

Text mining and analysis involves information retrieval, distributions, lexical examination to study word recurrence cognizance, labelling and annotation, data extraction, information mining methods including connection and affiliation investigation, visualization and prescient examination (Gupta & Malhotra, 2015). In other words, text mining and analytics helps in transforming explicit and documented content data into information for investigation. There are many features in text analytics tools that can extract and count the occurrence of terms in documents which is needed in the ontology development. (Analytics, 2017) has listed several text mining and analytics software along with the features and functionalities. Among the highlighted text mining and analytics software are QDA Miner Lite as a free computer assisted qualitative analysis software that can analyze textual data such as interview and news script, GATE is General Architecture for Text Engineering to process natural language and engineering language, TAMS Analyzer is for Macintosh OSX that can identify themes in texts such as website and interviews, Carrot2 does text and search result clustering framework, CAT as a free service of the Qualitative Data Analysis Program and RapidMiner as an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.

This research employs RapidMiner Text Analytics tools (Rapid Miner, 2017) for the financial criminal research documents text processing that able to provides data mining and machine learning procedures including Data Loading and Transformation (Extract, Transform and Load – ETL), data processing and visualization. It is accessible as a stand-alone application for information investigation and as a data mining engine for the integration into its own products. RapidMiner can support almost all of the Big Data Analysis and Data Science required processes such as Data Preparation, Modeling and Validation, Operationalization and Visualization. For this research, we require an addition Text Processing Extension that provides data and text mining features under the license of Educational Program. The details explanation on RapidMiner implementation on document processing of Financial Criminology research papers will be described further in the next section.

3. Methodology

The section describes the methodology in designing an ontology-based financial criminology knowledge representation data using text analysis process. The research process has been divided into three stages as below:

3.1 Sample Data (Journal/Research/Articles on Financial Criminology)

In order to capture and extract the commonly talked topics and terms in financial criminology, the research employs RapidMiner Text Analytics tools to process, extract and visualize financial criminology domain from twenty-five (25) research papers in the form of journals and articles in relation to ‘Financial Criminology’, ‘Financial Crime’ and ‘White-Collar Crime’ as the searching keywords. These research papers were downloaded from numerous and reputable online journal databases such as ProQuest, ScienceDirect, Scopus and Web of Sciences (WoS) with five to six papers from each database.

3.2 Text Mining and Analysis Implementation with RapidMiner Text Processing Extension

The process of data mining and text processing in RapidMiner will go through several machine learning procedures which are Process Document From Files, Tokenize, Transform Cases, Filter Stop words, Generate n-Grams and Filter Token (By Length). Figure 1 shows the overall process in Text Mining procedure with RapidMiner for this research.

Fig. 1 Text Mining and Analysis Process in RapidMiner

3.2.1 Process Document Files

RapidMiner Text Processing requires creating a Process which is the root operator of every analysis process. In this case, the research need to create a Process called Process Document from Files that allows us to specify a number of text collection stored in multiple files – the twenty-five (25) financial criminology research papers. As shown in figure 2, directory containing research papers document in .pdf format is selected for processing.

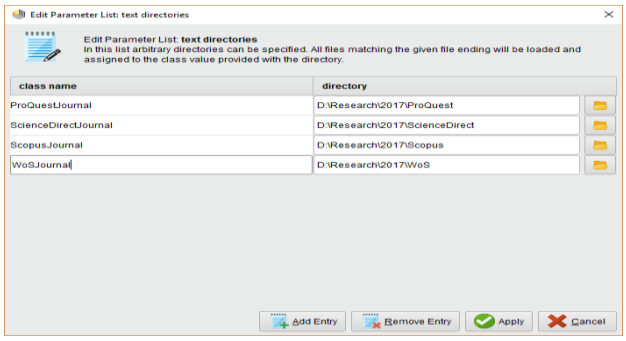


Fig. 2 Process Document Operator and Document Selection

3.2.2 Tokenization

Tokenize operator will splits the text of a document into a sequence of tokens. There are several options or mode of splitting points such as to use all non-letter characters, specify characters, regular expression, linguistics sentences and linguistics tokens. In this research, we chose all non-letter characters which will extract tokens consisting of one single word. This is the most appropriate option before finally building the word vector (RapidMiner, 2017).

3.2.3 Transform Cases

The next operator is Transform Case which is to transform all characters in a document to either lower case or upper case respectively. In this case, we transform all characters into lower case.

3.2.4 Filter Stopwords

In order to make sure the text processing be able to extract commonly term in the domain correctly, the research need to remove most of the English Stop Words such as {a, about, above, also, all, almost, alone... n}. The Filter Stop Words Operator did its job by removing every token which equals a stop words from the RapidMiner built-in stop words list.

3.2.5 Generate n-Grams

The next process is to generate n-Grams that create term n-Grams of tokens in a document. This operator creates term n-Grams of tokens in a document. A term n-Gram is defined as a series of consecutive tokens of length n. The term n-Grams generated by this operator consist of all series of consecutive tokens of length n (RapidMiner, 2017). In this research, the max length of n is 2.

3.2.6 Filter Token by Length

Finally, the research uses a Filter Token (By Length) to filter token based on the length of characters. The parameter for this operator is Minimum Characters and Maximum Characters. In this research, we chose the minimum characters of every token to 4 characters and the maximum of 25 characters. Only token or term extracted within this range will be considered in the ontology development.

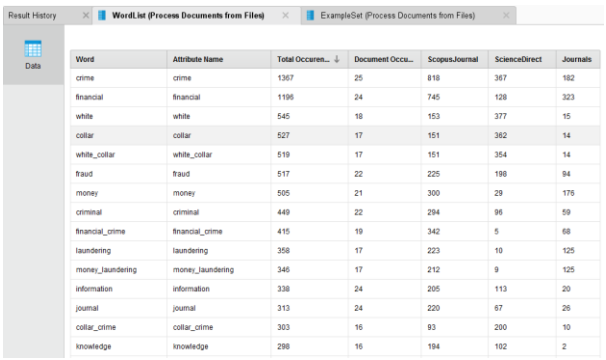
3.3 Ontology Development

Based on the result of text mining and processing by RapidMiner, the research continues with the development of Ontology design model. The process of ontology development involves four main steps as suggested by Noy & McGuinnes (2001), which are enumerating all important terms in the domain, identifying terms' definition and related meaning, defining classes and class hierarchy and finally identifying relationship between classes. Term enumeration process is performed to understand what is the commonly shared knowledge used in all data sources. This process will be done by RapidMiner Text Processing tools with defined machine learning procedures. The class is identified by referring to the meaning of all terms enumerated. The most generalize term that has meaning and purpose which can represent a specific category is chosen as the class name. Then, the research classifies and verify the instances of the categories into it classes or concepts identified previously by consulting with the expert. Thesaurus and dictionary also was used to understand the meaning of each term and its concepts. This research applied the top-down process where class identification starts from general to specialize class. The next step is the implementation stage which is the process to transform the ontology design model into preferred knowledge representation language, Ontology Web Language (OWL). OWL can be best developed by using the ontology development semantic editor called Protégé version 5.2.0. All classes, subclasses, relationship and instances which have been identified in previous section have been transformed as ontology visualization by using the Protégé ontology development editor. The step by step practical guideline in developing ontology by using Protégé 5.2.0 was referred to Prot et al., (2011).

4. Result and Discussion

This section describes the result of this research. Based on the text mining and analysis by RapidMiner Text Processing tools towards twenty-five (25) resource documents related to journal and articles on Financial Crime, Financial Criminology and White-Collar Crime, the result can be shows as figure 3.

The result shows all tokens or words along with number of occurrence for each word. The total number of words extracted was 880 words. The words extracted based on the machine learning procedures configured in our methodology discussed in previous section. There are no stop words counted, the minimum length for each word is four characters and maximum is twenty five characters. The result indicates the highest number of words counts occurrence, words and attribute names from each document. The highest number of words occurrence are 'crime' with 1367 occurrence, 'financial' with 1196 total occurrence, 'white-collar' with 519, 'fraud' with 517 and so forth. Based on this result, the research successfully enumerate all the commonly terms and shared knowledge used in all data sources as per required in the ontology development process. Later, the terms defined by the text processing is used to identify the concepts or classes, attributes or relationship and instances of the domain.



Word	Attribute Name	Total Occuren...	Document Occu...	ScopusJournal	ScienceDirect	Journals
crime	crime	1367	25	818	367	182
financial	financial	1196	24	745	128	323
white	white	545	18	153	377	15
collar	collar	527	17	151	362	14
white_collar	white_collar	519	17	151	354	14
fraud	fraud	517	22	225	198	94
money	money	505	21	300	29	176
criminal	criminal	449	22	294	96	59
financial_crime	financial_crime	415	19	342	5	68
laundering	laundering	358	17	223	10	125
money_laundering	money_laundering	346	17	212	9	125
information	information	338	24	205	113	20
journal	journal	313	24	220	67	26
collar_crime	collar_crime	303	16	93	200	10
knowledge	knowledge	298	16	194	102	2

Fig. 3 Text Analysis Result for Words Extraction and Occurrence

In ontology model development process, the research found that there are nine main classes of concepts that are commonly talked and researched in the field of financial criminology. The classes are People, Offenses, Risk_Sector, Enforcement, Technology, Location, TimeFrame, Resources and Property. The defined classes and its relevance instances can be best described as below taxonomy.

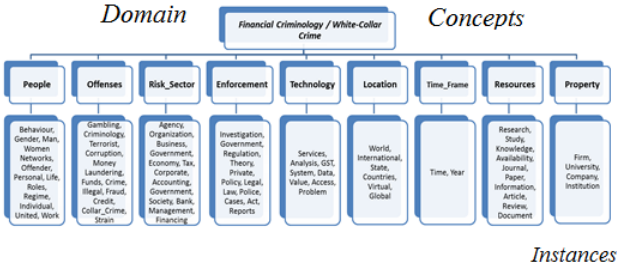


Fig. 4 Text Analysis Result for Words Extraction and Occurrence

Figure 4 describe the taxonomy of Financial Criminology found in this research. The taxonomy helps in organizing the result between classes and subclass or between classes and instances in the ontology. This related to the relationship between concepts or classes through hierarchical and associative relationships. Hierarchical relation is the relationship between classes and instances within the same hierarchy, meanwhile associative relationship define the relation between class and instance in different concept hierarchy.

Based on the taxonomy, the research continues to implement the ontology (Classes/Concepts, Relationship, Instances) into preferred knowledge representation language, Ontology Web Language (OWL). The OWL is a family of Knowledge Representation languages for authoring Ontologies. The result shows that the defined ontology classes, attributes and instances are able to transform in ontology OWL. The ontology representation of Financial Criminology domain classes is shown in figure 5. The figure describe that financial criminology domain has nine main classes which are People which talked about individual or group that involve in financial criminology activity, Offenses class describe the types of crimes in financial criminology such as gambling, corruption and money laundering. The Risk Sector explains the possible occurrence of financial criminology in that sectors such as government sector, business sector, management and economy. Enforcement class explains the activity involves in the enforcement of financial criminology. Technology class commonly talked in financial criminology domain in the form of IT services, system, data value and analysis. Location and Time are classes of financial criminology which define the location and time coverage that financial crime normally occur whether in the state and country level, international and globally or even in virtual world. Resources class explains the sources of data and information that the researcher gained during the study of financial criminology and investigation. Finally, the Property class is generally explains on types of property that involves n financial criminology.

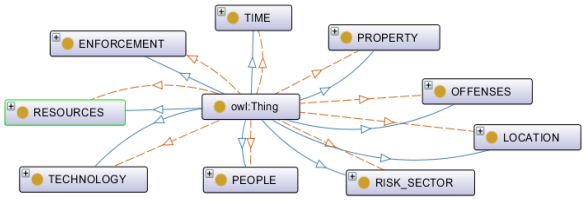


Fig. 5 Financial Criminology Domain Classes

Figure 6 describe the overall view of ontology representation of financial criminology for this research. It shows the relationship between the domains Financial Criminology noted as 'Thing' and its nine classes. The ontology view that uses the built in Ontology graph called OntoGraf also be able to show the objectProperty (Relationship types) noted as Has (Example : Thing Has Location) and its Data Property Instances.

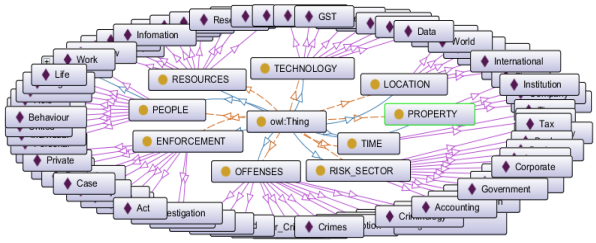


Fig. 6 The Ontology Representation of Financial Criminology

5. Conclusion

This research aims to represent and understand the financial criminology domain in the form of Ontology Knowledge Representation. The benefits of the domain representation are to helps researcher and investigator to understand the area of study and key issues in financial criminology. Meanwhile, Information Technology people are able to use the ontology (Semi or Structured Knowledge) to develop a knowledge base system of financial crime. In the process of data mining and text processing, the research employs a Big Data Analysis tools called RapidMiner that have additional text processing extension. The data extracted from 25 research papers on Financial Criminology gained from main online journal databases such as Scopus and Web of Knowledge (WoS). Then, the research uses the terms and words to design the ontology based on the methods explained in the methodology section. The result found that there are nine main classes or concepts are commonly research on financial criminology namely; People, Location, Time, Property, Offenses, Risk Sector, Enforcement, Technology and Resources. The ontology has successfully implemented in a Web Ontology Language (OWL) called Protégé 5.2.0 and the overall view of ontology graph that show the relationship between domain, classes and instances are presented in the result section.

Acknowledgments

We thank the Institute of Research Management and Innovation (IRMI) of Universiti Teknologi MARA (UiTM) for supporting this research through the Internal Research Acculturation Grant Scheme (iRAGS) with project code number 201215160004.

References

[1] Analytics, P. (2017). Top 27 Free Software for Text Analysis, Text Mining, Text Analytics - Predictive Analytics Today. Retrieved February 10, 2017, from https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

[2] Brachman, R., & Levesque, H. (2004). Knowledge Representation and Reasoning. The Morgan Kaufmann Series in Artificial Intelligence (Vol. 1). http://doi.org/10.1146/annurev.cs.01.060186.001351

[3] Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8(2), 240–247. http://doi.org/10.1016/S0022-5371(69)80069-1

[4] Fridson, M. S. (2002). Financial Crime Investigation and Control. Financial Analysts Journal, 58(5), 99–100. Retrieved from http://search.proquest.com/docview/219247914?accountid=14549%5Cnhttp://hl5yy6xn2p.search.serialssolutions.com/?genre=article&sid=ProQ:&atitle=Financial+Crime+Investigation+and+Control&title=Financial+Analysts+Journal&issn=0015198X&date=2002-09-01&volume=5

[5] Gottschalk, P. (2010). Categories of financial crime. Journal of Financial Crime, 17(4), 441–458. http://doi.org/10.1108/13590791011082797

[6] Grenier, L. (1998). Working with indigenous knowledge: A guide for researchers. Retrieved from https://books.google.com/books?hl=en&lr=&id=VlGncLGEf5MC&oi=fnd&pg=PR5&dq=Working+With+Indigenous+Knowledge&ots=gqwmGY9GKD&sig=fv2l2FYTfobOrLl58PL0rGhfzMM

[7] Gupta, G., & Malhotra, S. (2015). Text documents tokenization for word frequency count using RapidMiner (Taking resume as an example). International Journal of Computer. Retrieved from https://pdfs.semanticscholar.org/97d0/d78e7d26141defdc917cad8307291c6f7e27.pdf

[8] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems. http://doi.org/10.1016/j.dss.2010.08.006

[9] Nour, P. (2003). Ontology-based retrieval of software engineering experiences. Retrieved from http://pages.cpsc.ucalgary.ca/~maurer/uploads/Publications/Nour2003.pdf

[10] Noy, N. F., & McGuinness, D. L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology.

Stanford Knowledge Systems Laboratory, 25. http://doi.org/10.1016/j.artmed.2004.01.014

[11] Obomsawin, R. (2001). Indigenous knowledge and sustainable development. Retrieved from https://scholar.google.com/scholar?q=Indigenous+knowledge+and+sustainable+development.+Obomsawin&btnG=&hl=en&as_sdt=0%2C5

[12] Prot, U., Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C., … Brandt, S. (2011). A Practical Guide To Building OWL Ontologies Using Protégè 4 and CO-ODE Tools Edition 1.3. Matrix, 0–107. Retrieved from http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf

[13] Rapid Miner. (2017). Data Science Platform | RapidMiner. Retrieved February 13, 2017, from https://rapidminer.com/

[14] Rosenberg, J. (1986). Dictionary of artificial intelligence and robotics. Retrieved from https://www.osti.gov/scitech/biblio/5434585

[15] Yuanyuan, W. Y. W., Rujing, W. R. W., Xue, W. X. W., & Yimin, H. Y. H. (2010). Ontology-Based Knowledge Representation for Agricultural Intelligent Information Systems. Management and Service Science MASS 2010 International Conference on, 9–12. http://doi.org/10.1109/ICMSS.2010.5577004

[16] RC Chakraborty, 2010. Knowledge Representation: AI Course Lecture 15-22. Retrieved December, 2012, from www.myreaders.info/html/aritificial_intelligence.html

[17] Tanwar, P., Prasad, T. V., & Datta, D. K. (2010). An Effective Knowledge base system Architecture and issues in representation techniques. International Journal of Advancements in Technology http://ijict. org/ISSN, 0976-4860.

ZULAZEZE SAHRI received the B.S in Information Technology degrees from International Islamic University Malaysia in 2007 and M.S in Information Technology degrees from Universiti Teknologi MARA Malaysia in 2012. During 2007-2009, he worked in an Indian Based software house, Mahindra Satyam working on Business Management Process (BPM) based software development by using Global 360 BMP tool and 2009-2010 he worked with a Computer Security Outsourcing Company, Digicert Sdn Bhd which gave him experienced in setting up a Security Operation Center for clients. He is now with the University Teknologi MARA (UiTM) as an academic staff. His research interest is in e-Commerce Development for Small Medium Enterprise in Malaysia and closely working with the clients to study for the enhancement of sales and purchase using e-Commerce Analytics. He has a few number of clients' accounts in managing Search Engine Optimization (SEO) and Search Engine Marketing (SEM) as part of his research and consultation.