International Journal of Academic Research in Business and Social Sciences
2017, Vol. 7, No. 12
ISSN: 2222-6990

Ontology Population from Textual Document Sources for Environmental Management Domain based Lexical Patterns Technique

Zaharudin Ibrahim, Tengku Adil Tengku Izhar, Mohd Sazili Shahibi, Mohd Ridwan Seman@ Kamarulzaman and Ahmad Zam Hariro
Faculty of Information Management, Universiti Teknologi MARA, UiTM, Selangor, Malaysia

DOI: 10.6007/IJARBSS/v7-i12/3729 URL: http://dx.doi.org/10.6007/IJARBSS/v7-i12/3729

Abstract
This study focuses on the approach of identifying semantic relationships from unstructured textual documents related to river water pollution from websites and proposes a lexical pattern technique to acquire the instances. This study has identified 10 types of concepts (entities), 10 object properties (or semantic relations) and twenty lexico-syntactic patterns have been identified manually, including one from the Hearst hyponym rules. The lexical patterns have linked 45 terms that have the potential as instances. Based on this study, it is believed that determining the lexical pattern at an early stage is helpful in selecting relevant term from a wide collection of terms from the corpus. However, the relations and lexico-syntactic patterns or rules have to be verified by domain expert before employing the rules to the wider collection in an attempt to find more possible rules. This study shows that background knowledge about the domain is essential to develop the TBox ontology diagram that serve as backbone of the domain ontology. This diagram is essential as guideline in discovering lexico-syntactic patterns therefore expedite the knowledge extraction process.
Keyword: Environmental Management, Lexical Patterns Technique, Ontology, Textual Document

1. Introduction

First introduced by Aristotle, ontology has recently become a topic of interest in computer science. Ontology provides a shared understanding of the domain of interest to support communication among human and computer agents; it is typically represented in a machine processable representation language (Haase and Sure, 2004) and is also an explicit formal specification of terms, which represents the intended meaning of concepts, in the domain and relations among them, and considered as a crucial factor for the success of many knowledge-based applications (Staab, et al, 2001). With the overwhelming increase in biomedical literature in digital forms there is a need to extract knowledge from the literature **(**Fuller, et al, 2004). Ontology may also be helpful in fulfilling the need to uncover information present in large and unstructured bodies of text, commonly referred to as non-interactive literatures (Swanson & Smalheiser, 1997) i.e., literatures that do not cite each other but which, nevertheless, together present useful new information.

International Journal of Academic Research in Business and Social Sciences
2017, Vol. 7, No. 12
ISSN: 2222-6990

Ontology is considered as the backbone of many current applications, such as knowledge-based systems, knowledge management systems and semantic web applications. One of the important tasks in the development of such systems is knowledge acquisition. Conventional approaches to knowledge acquisition are mainly from interviewing domain experts and subsequently modelling and transforming the acquired knowledge into some form of knowledge representation technique. However, a huge amount of knowledge is currently embedded in various academic literatures and has the potential of being exploited for knowledge construction. The main inherent issue is that such knowledge is highly unstructured and difficult to transform into meaningful model. Although a number of automated approach in acquiring such knowledge has been proposed by Alani, et al (2003) and Cimiano, et al (2005) their success have yet to be seen. Such approaches have only been tested on general domain and scientific domains such as the medicinal herbs domain have yet to be explored. While automated approach seems to offer promising solutions, human still play an important role in validating the correctness of the acquired knowledge, particularly in scientific domain. This study, therefore, proposed a semi-automated approach for discovering domain-specific concepts and relationships in an attempt of acquiring domain knowledge of the medicinal herbs domain from web documents. The Hearst's technique (Hearst, 1992) has been employed to extract concept terms from the literature and to discover new patterns through corpus exploration. The technique acquires hyponym relations automatically by identifying a set of frequently used unambiguous lexico-syntactic patterns in the form of regular expressions. The study of lexico-syntactic pattern also was carried out by Moldovan, et al (2000) which discovered domain-specific concepts and relationships in an attempt to extend the WordNet. The pattern-based approach was also applied by Pantel & Pennacchiotti (2006) which proposed a bootstrapping algorithm to detect new patterns in each iteration. There are also studies by Imsombut & Kawtrakul (2007) that proposed methods which are very close to the pattern-based approach of extracting the ontology from item lists, especially in technical documents, and by Zaharudin, et al (2009) which applied a semi-automatic approach in collecting relevant terms extracted from lexico-patterns from medicinal herb documents to be inserted as ontological term.

The main objective of this study is to explore the possibility of identifying semantic relationships (SR) between terms from environmental management domain literature and to represent it in the form of some ontological structure. Since ontology can be represented in many forms, ranging from a simple list of concepts to the complexity of logic structure (Zaharudin, et al, 2009), this study focuses on representing the concepts that ismanually selected by expert with semantic links and discover the lexico-syntactic patterns or semantic rules. The rules created from analyzing a set of textual documents will be used to populate medicinal herb domain ontology. The paper focused on our initial findings based upon current approaches of extracting knowledge from unstructured documents as previously described.

2. Materials and Methods

The advances in the biomedical sciences need the development of ontology to help user understand the developments in one's own area of specialization, and also to enable them to

quickly learn about the developments in related and unrelated subject areas. Ontology plays an important role in facilitating the formal sharing and re-using of knowledge through the construction of an explicit domain model (Fensel, 2004; Gruber, 1995). However, the manifestation of this role requires the construction of concepts in the particular domain. The meaning of terms and the relationships between entities within that particular domain of knowledge must be well defined. This is to ensure that the knowledge can be interpreted logically before a body of knowledge can be communicated unambiguously across computer-based systems (Catton, et al, 2003).

2.1 Rules for Populating Environmental Domain Ontology

Our approach mainly concerns with identifying entities related to the domain such as ’agriculture’, ’urban area’, ‘pollution’ as shown in Fig. 1.

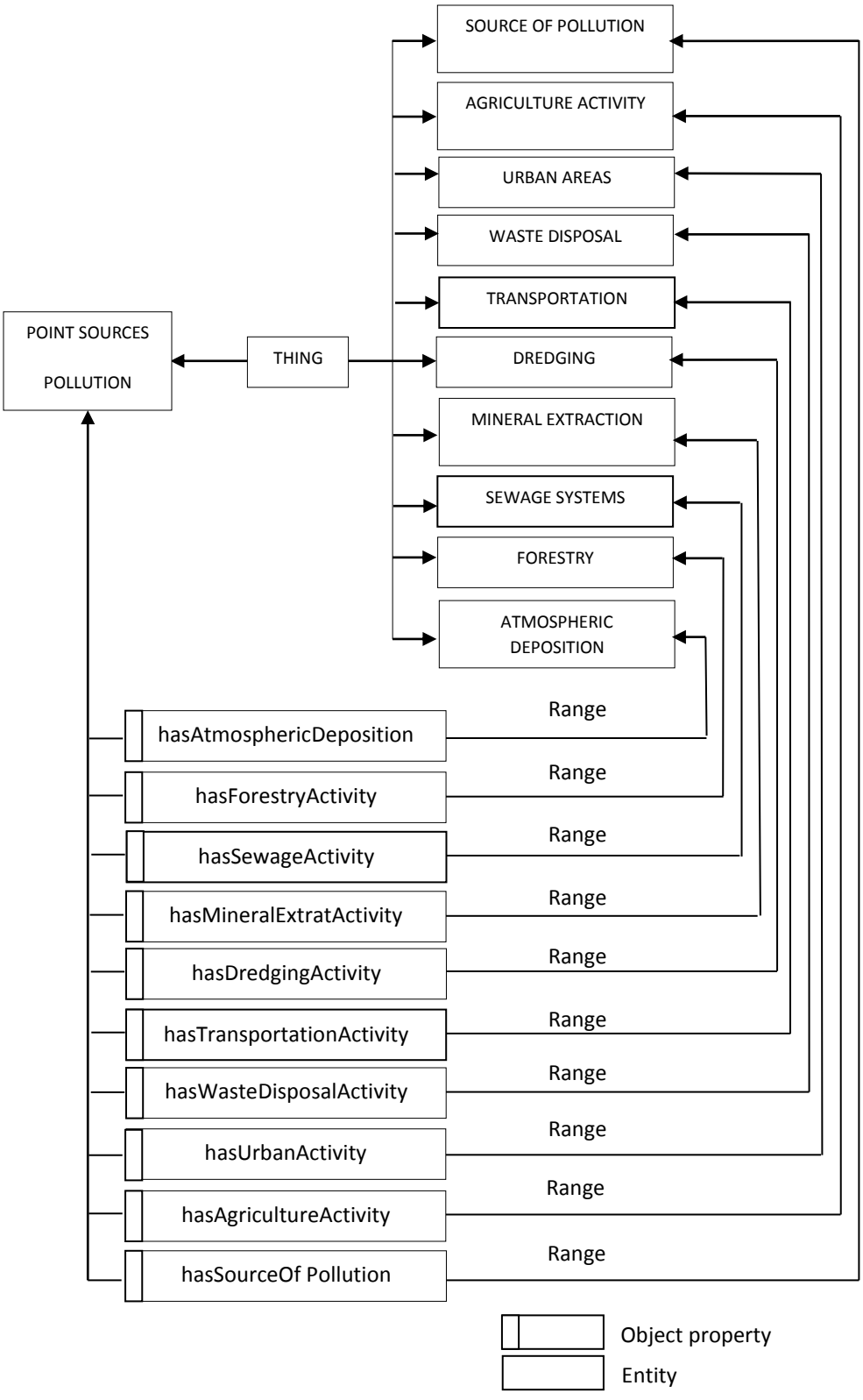
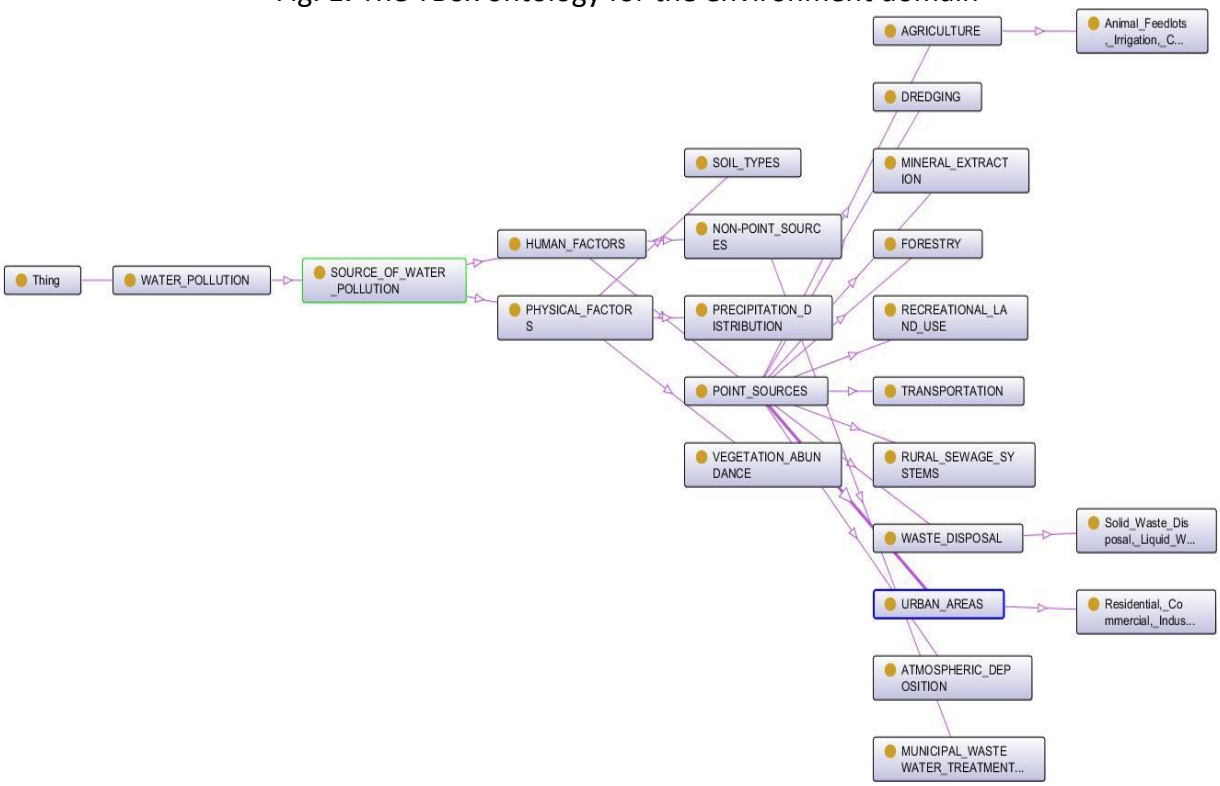

SOURCE OF POLLUTION
AGRICULTURE ACTIVITY
URBAN AREAS
WASTE DISPOSAL
TRANSPORTATION
DREDGING
MINERAL EXTRACTION
SEWAGE SYSTEMS
FORESTRY
ATMOSPHERIC DEPOSITION
POINT SOURCES POLLUTION
THING
hasAtmosphericDeposition
hasForestryActivity
hasSewageActivity
hasMineralExtratActivity
hasDredgingActivity
hasTransportationActivity
hasWasteDisposalActivity
hasUrbanActivity
hasAgricultureActivity
hasSourceOf Pollution
Range
Object property
Entity

Fig. 1. The TBox ontology for the environment domain



As can be seen from Fig. 1 there are ten concepts (entities) and ten object properties (or semantic relations) in the constructed ontology. The rules designed for name entities recognition are mainly set of patterns which matched with the given object properties and subsequently inferred to be instances of the corresponding concepts. These ten semantic relations are constructed according to stages illustrated in Fig. 2.

Stage 1: Collect documents from websites
Stage 2: Apply Hearst's rules on hyponyms identification
Stage 3: Discover new concepts
Stage 4: Extract sentences
Stage 5: Discover lexico-syntactic patterns
Stage 6: Extract new concepts.

Fig. 2. The stages of the semantic relation identification

For the initial stage, several documents related to river water pollution have been collected from websites. We then apply Hearst's rules to identify hyponyms. In this case the collection or corpus was analyzed to extract the semantic relationships enhancement based on the work of Hearst (1992). The Hearst rule for detecting hyponym from text includes:

1. NPo . . .such as{NP1,NP2 . . .(and\or)}NP, ,
2. such NP as{NP,} _ {(or [and)}NP
3. NP{,NP} _ {, } or other NP
4. NP{,NP} _ {, } and other NP
5. NP{, } including {NP _ {or\and} NP
6. NP{, } especially {NP,} _ {or] and}NP

3. Research Approach

In this study, a semi-automatic approach using NLP tools was used to collect relevant terms in medicinal herb documents to be inserted as ontological term. This approach requires that the researchers be familiar with the terms or concepts in the domain. They should have substantial knowledge regarding the domain before selecting the appropriate terms that can be linked to each other.

Stage 1: Document Collection and Preparation.

For the initial stage as shown in Fig. 2, several documents related to medicinal herbs have been collected from websites.

Stage 2: Apply Hearst's rules on hyponyms identification

The collection was analysed to extract the semantic relationships enhancement based on the work of Hearst (1992) as described in previous section. In order to find other semantic relation rules or lexico-syntactic patterns, the technique of seed words from Moldovan, et al (2000) was modified as described below.

Stage 3: Discover new concepts

In this study a few terms have been selected as seed concepts that are considered important. The focus of this study is on the environmental domain domain and the words selected are *agricultural.*

Stage 4: Extract sentences

The documents retrieved were further processed so that only sentences containing the seed words were retained. Each sentence in this corpus was part-of-speech (POS) tagged and then parsed using Genia Tagger.

Stage 5: Discover lexico-syntactic patterns

The approach is to search for lexico-syntactic patterns comprising the concepts of interest.

Stage 6: Extract new concepts.
After discovering lexico-syntactic patterns, the new concepts which are directly related to the seeds were extracted from the sentence. In the example above, the seed word (Maca) was linked with new concepts such as libido and fertility, erectile dysfunction and ED.The analysis processes were simplified into the following stages as shown in Fig.2.

4. Results and Discussion

In this study, several documents from the websites relating to medicinal herbs have been selected for corpus development. Even though Hearst technique intensively applied to linguistic domain but in this the rules was modified to accommodate the purpose of the medicinal herbs domain. The modification involved combination of Hearst techniques and the seed words technique used by Moldovan. The Hearst rule for detecting hyponym from text includes:

(l) *NPo ..... such as {NP1, NP2 . . . . (and /or)} NP,,*
(2) *such NP as {NP ,}* {(or [and)} NP*
(3) *NP {, NP} * {,}* or *other NP*
(4) *NP {, NP}* {,} and other NP*
(5) *NP* {,} *including {NP* * {or / and} NP
(6) *NP* {,} *especially {NP,}* {or] and} NP*

The identification of other relations was heuristically selected if there is a relation between a pair of noun phrase appearing in each of the sentence. The lexical patterns were identified from the relationship between terms (noun phrase) in a sentence. In this study only hyponym has a proper identification approach. Other relations in this list will be investigated further to determine the appropriate relations. This study initially found seven types of concepts (entities) and eight object properties (or semantic relations) in the constructed ontology. The entities are: EFFECT, FORM OF, FUNCTION, HARMFUL, USAGE, PROMOTE and HERB as shown in Figure 1.

In this study, eight types of semantic relations were found. From these relations, twenty new lexico-syntactic rules were derived for semantic relationships. The details of the rules are given in Table 1.

Table 1: Types of relation terms and rules for semantic relation or lexical pattern

ENTITY	LEXICAL-PATTERN	TEXT	SEMANTIC RELATION	SOURCE
SOURCE	The….source of pollution in/on NP_0 is/are NP1, Np2, Np3 and/or NP HasPollutionSources(NP_0, NPi)	The domestic pollution is the major source of pollution in Gang a river.	HasPollutionSources(NP_o,NPi)	http://shodhganga.inflibnet.ac.in/bitstream/10603/23761/6/06_chapter%201.pdf
		"Stormwater runoff is the number one source of pollution in the Charles River,"...		www.wbur.org/cognoscenti/2016/08/26/boston-climate-change-frederick-hewett
		A key non-point source of pollution on the St. Johns River is stormwater runoff from farms, ...		http://www.theriverreturns.org/protect/solutions/
	The NP_0 is polluted.. due to .. NP1, Np2, Np3 and/or NP HasPollutionCause(NP_0, , NP_i)	On the basis of this study it was concluded that water of Tapi River is moderately polluted due to discharges of industrial waste, domestic sewage and agricultural run-off in river water ...		http://www.academia.edu/4772160/water_quality_and_pollution_status_of_tapi_river_gujarat_india
		The Yangtze river has become severely polluted due to industrial development…		http://www.greenpeace.org/international/en/multimedia/slideshows/Fishing-and-Pollution-in-Yangtze-River-China/
		Indian rivers are p		https://unstats.un.org/unsd/

		olluted due to the dis- charge of untreated sewage and industrial effluents.		environment/envpdf/pap_wasess5a2india.pdf
	The NP_0 is contaminated by {NP1,}* {or/and} {NPi}	The Ok Tedi River is contaminated by tailings from a nearby mine.	HasContamination(NP_o,NPi)	https://en.wikipedia.org/wiki/Environmental_impact_of_mining
	HasContamination(NP_0,NPi)	the Roseau River, from the Bath Estate Bridge to the sea, because that section of the Roseau River is contaminated by sewage and other pollutants.		http://dominicanewsonline.com/news/homepage/news/health/moh-warns-of-roseau-river-contamination/
		Rooiwal River is contaminated by sewage..		http://rekordnorth.co.za/88259/rooiwal-river-time-bomb/
		The Snake River is contaminated by acid rock-drai-nage,		http://www.umass.edu/biocomplexity/index2_files/AMD16.pdf
		…Aberjona River is contaminated by volatileorganic compounds, ...		https://www.atsdr.cdc.gov/hac/pha/pha.asp?docid=1285&pg=5Ground
		... The Negeram River is contaminated by waste material ….		http://scialert.net/fulltext/?doi=jas.2012.1801.1808&org=11
		Oustouan River is contaminated by waste water,		https://www.omicsonline.com/open-access/study-of-physical-and-chemical-parameters-of-oustouan-river-north-lebanon-2473-3350-1000430.php?aid=76712&view=mobileThe

Agricultural activities	Agricultural activities * can lead to* {NP1,}* {or/and} {NPi} into {NP_0}	Agricultural activities can lead to an increased influx of nitrogen into Maines River.	HasAgriculturalActitvityEffect(NP_0, NPi)	http://www.fao.org/docrep/004/y3618e/y3618e07.htm
	Agricultural activities * a source of * {NP1,}* {or/and} {NPi} into { NP_0}	Agricultural activities may be a source of pesticide pollution of Danube River	HasAgriculturalActitvityEffect(NP_0, NPi)	
Urban activities		….. urban activities have substantially affected water quality in several areas of the South Platte River Basin, investigation of water quality by the U.S. Geological Survey.	HasUrbanActitvityEffect(NP_0, NPi)	https://archive.usgs.gov/archive/sites/www.usgs.gov/newsroom/article.asp-ID=1136.html
		Industrial and urban activities in the Leeds catchment of the River Aire have caused the widespread dispersal and contamination by a variety of ...		etheses.whiterose.ac.uk/2662/
		various industrial and urban activities which adversely affect the river .		shodhganga.inflibnet.ac.in/jspui/bitstream/10603/114482/11/11_chapter%201.pdf
waste disposal		Every day the Yamuna river is	HasWasteDisposal(NP_0, NPi)	file.scirp.org/Html/12-9401064_1806.htm

		polluted by solid waste disposal, animal bathing, disposal ...		
		The municipal waste disposal vats near the river also causing water pollution.		www.iosrjournals.org/iosr-jestft/papers/vol10-issue9/Version-2/C1009021925.pdf
forestry activities		The most common form of stream pollution associated with forestry activities	HasForestryActivity(NP_0, NPi)	http://www.who.int/water_sanitation_health/resourcesquality/watpolcontrol.pdf
transportation activities		transportation activities, such as bridge construction, repair, replacement, or retrofitting, and .... known projects will affect the Virgin River or its habitats in Utah.	HasTransportationActivity(NP_0, NPi)	
Dredging activities		Dredging in southern Vietnam *river* leads to serious *pollution* .	HasDredgingActivity(NP_0, NPi)	tuoitrenews.vn/society/.../dredging-in-southern-vietnam-river-leads-to-serious-polluti...
	Dredging cause of {NP1,}* {or/and} on/along { NP}	.. dredging cause of landslides along Dong Nai River.		
	Dredging will affect a NP0's NP1, NP2 or/and NPn.	Dredging will affect a river's composition, diversity and resiliency in a		https://www.vietnambreakingnews.com › News www.thedailystar.net › Opinion › Environment

		variety of ways.		
Mineral extraction	Mineral extraction activities, including*	Mineral extraction activities, including coal mining, conventional and unconventional oil and gas development, and waste disposal have historically and continue to exert pressure on water quantity, water quality, and ecological health in the upper Ohio River Drainage Basin.	Has Mineral extractionActivity(NP_0, NPi)	http://www.lrp.usace.army.mil/Portals/72/docs/WaterManagement/Water%20Quality%20Concerns.pdf
sewage water	……suffering from NP1, NP2… due to… the release of sewage water into NP0	People in downstream villages are suffering from various health-related problems due to the release of sewage water into Malaprabha river	HasSewageAffect(NP,NPi)	http://timesofindia.indiatimes.com/city/hubballi/Sewage-water-released-into-Malaprabha-river/articleshow/55006790.cms
	- NP1 gallons/litres of sewage spilled into NP0	- About 143 million gallons of sewage spilled into the Tijuana River		www.sandiegouniontribune.com/news/.../sd-me-sewage-spill-20170224-story.html
	… after pumping NP1…. sewage into NP0	Thames Water has been fined a record £20m after pumping 1.9 billion litres of untreated sewage		http://www.bbc.com/news/uk-england-39352755

		into the River Thames		
	…gallons of …sewage dumped Into NP0	100,000 Gallons of Raw Sewage Dumped Into Hudson River		http://wpdh.com/100000-gallons-of-raw-sewage-dumped-into-hudson-river/
	…. Sewage* released into NP0 .. implications for Np1, Np2..	Untreated or poorly treated sewage is being released into rivers and storm drains with serious implications for human health and the natural environment.		http://www.scsoft.de/et/et2.nsf/KAP2View/717882FEC1CDECC605256760006AAD33?OpenDocument
	The threats to NP1, NP2.. sewage pollution of NP0	The threats to human health from sewage pollution of rivers		http://www.scsoft.de/et/et2.nsf/KAP2View/717882FEC1CDECC605256760006AAD33?OpenDocument
ATMOSPHERIC DEPOSITION	Atmospheric deposition of NP1 in the NP0	Atmospheric deposition of nitrogen in the Mississippi River Basin	HasAtmosphericDeposition(NP,NPi)	https://toxics.usgs.gov/pubs/wri99-4018/Volume2/sectionC/2413_Lawrence/pdf/2413_Lawrence.pdf
		atmospheric deposition of PCBs to the river		http://pubs.acs.org/doi/pdf/10.1021/es052149m
		Direct and indirect atmospheric deposition of PCBs to the Delaware River watershed.		https://www.ncbi.nlm.nih.gov/pubmed/16646449

In this preliminary stage, the ontology populating process has been done manually by mapping all the rules to the documents. As the result about fifty one (51) instances have being extracted as shown in Table 2.

Table 2. List of terms extracted using lexical pattern.

List of terms		
1.9 billion	Health-related problems	Retrofitting
143 million	Human health	Serious pollution
Acid rock-drainage,	Human health	Sewage
Agricultural run-off	Increased influx of nitrogen	Solid waste
Animal bathing	Industrial development	Storm water runoff
Bridge construction	Industrial effluents	Stream pollution
Coal mining	Landslides	Tailings from a nearby mine.
Composition	Natural environment	Untreated sewage
Contamination	Nitrogen	Volatile organic compounds
Discharges of industrial waste	Oil and gas development	Waste disposal
Disposal	PCBs	Waste material
Diversity	Pesticide pollution	Waste water
Domestic pollution	Pollutants	Water pollution
Domestic sewage	Repair	Water quality
	Replacement	Widespread dispersal
	Resiliency	

By using the approach proposed in this study all instances that linked by the lexico-syntactic rules have been considered as ontological terms. The knowledge extraction process whereby the instances were discovered is illustrated by ABox diagram as shown in Fig. 3.

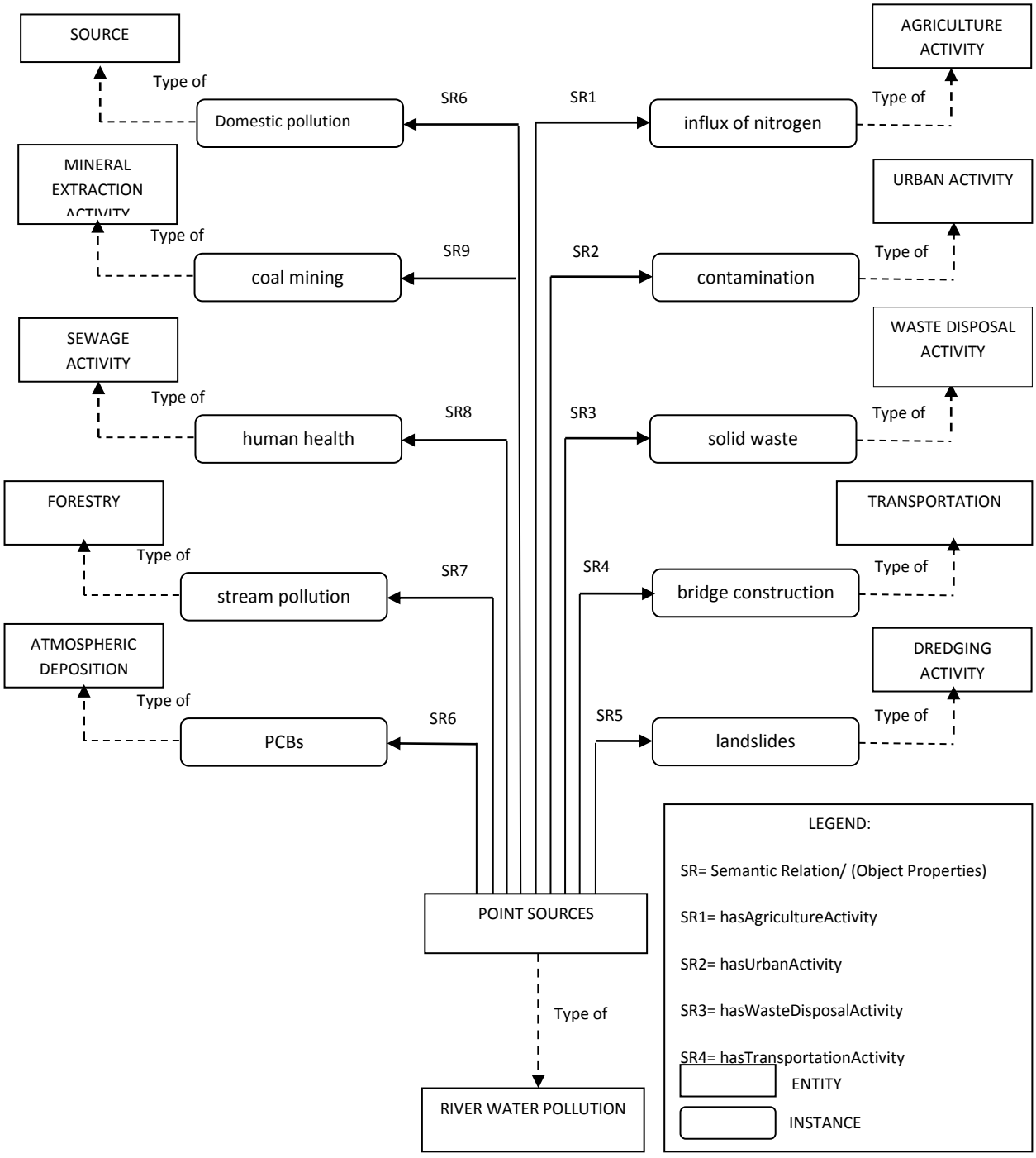


Fig. 3. The ABox Ontology Diagram

5. Conclusion

In this paper an approach of extracting semantic relation from a specific domain has been presented. This study emphasized on finding the lexical patterns at earlier stage after selecting the relevant terms. This study shows that it is helpful to have background knowledge about the domain prior to finding the lexical pattern. It is hoped that applying all the twenty nine semantic relation rules to other documents in the corpus will reduce the effort to eliminate irrelevant terms which previously has to be done manually. Therefore this approach will minimize the use of domain expert which sometimes is a constraint in the development of the specific domain ontology.

Acknowledgement

The authors would like to extend our deepest gratitude to the Universiti Teknologi MARA (UiTM) for financing this project under the LESTARI RESEARCH GRANT (Reference 600-IRMI/MYRA 5/3/LESTARI). Our thanks are also dedicated to Research Management Centre (RMC), Universiti Teknologi MARA (UiTM) for facilitating us towards the completion of the project. We would like to dedicate our thanks to Assoc. Prof. Dr. Kadir Arifin from the Faculty of Social Sciences and Humanities, National University of Malaysia.

References

Alani, H., S. Kim, D. E. Millard., M. J. Weal.,W. Hall., P. H. Lewis & N. R. Shadbolt, 2003. Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE. Intelligent Systems. 18(1): 14-21.
http://www2.computer.org/portal/web/csdl/doi/10.1109/MIS.2003.1179189

Catton, C., R. Dalton, C. Wilson, and Shotton (2003). Sabo: A Proposed Standard Animal Behaviour Ontology. Image Bioinformatics Laboratory, Department Of Zoology, University Of Oxford, UK**.** www.bioimage.org/pub/SABO/SABO_cornell_final.pdf.

Cimiano, P., A. Hotho & S. Staab, 2005. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research.24: 305-339. http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume24/cimiano05a.pdf

Fensel, D. (2004). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Berlin: Springer-Verlag Berlin Heidelbeg New York. ISBN: 3-540-00302-9

Fuller, S., D. Revere, P. Bugni and G.M. Martin**,** 2004. A knowledgebase system to enhance scientific discovery: Telemakus. *Biomedical Digital Libraries.* **1:**2. http://www.bio-diglib.com/content/1/1/2

Tagge. G.. http://text0.mib.man.ac.uk/software/geniatagger/ index. Html

International Journal of Academic Research in Business and Social Sciences
2017, Vol. 7, No. 12
ISSN: 2222-6990

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies. 43(5-6): 907 - 928. ISSN: 1071-5819. http://portal.acm.org/citation.cfm?id=219701

Haase, P & Y. Sure, 2004. State-of-the-Art on Ontology Evolution. Institute AIFB, University of Karlsruhe. http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.1.b.pdf

Hearst, M., 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics. 539-545. http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf

Zaharudin, I., S.A. Noah and M.M. Noor (2009). Knowledge Acquisition from Textual Documents for the Construction of Medicinal herb Ontology Domain. J. Applied Science. 9(4):794-798. http://www.scialert.net/qredirect.php?doi=jas.2009.794.798&linkid=pdf

Imsombut, A & A, Kawtrakul (2007). Automatic building of an ontology on the basis of text corpora in Thai. Journal of Language Resources and Evaluation. 42(2), 137-149. http://www.springerlink.com/content/h24p4p55882704r0/

Moldovan, D., R. Girju, and V. Rus (2000). Domain-specific knowledge acquisition from text. Proceedings of the sixth conference on Applied natural language processing. Seattle, Washington, 268 – 275. http://portal.acm.org/citation.cfm?id=974147.974184

Pantel, P & M. Pennacchiotti, (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics Sydney, Australia. 113-120. http://portal.acm.org/citation.cfm?id=1220175.1220190
Staab, S., H.-P. Schnurr., R. Studer & Y. Sure, 2001. Knowledge processes and ontologies: IEEE. Intelligent Systems, Special Issue on Knowledge Management. 26-34 . http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00912382

Swanson, D.R. & N.R. Smalheiser, 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence. 91(2): 183-203. http://portal.acm.org/citation.cfm?id=251766&dl=