9-2008

Collecting Open Source Intelligence via Tailored Information Delivery Systems

William Sousan
University of Nebraska at Omaha, wsousan@gmav.unomaha.edu

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Ryan Nickell
University of Nebraska at Omaha, rnickell@gmav.unomaha.edu

William Mahoney
University of Nebraska at Omaha, wmahoney@unomaha.edu

Peter Hospodka
University of Nebraska at Omaha

Collecting Open Source Intelligence via Tailored Information Delivery Services

W. L. Sousan, Q. Zhu, R. Nickell, W. Mahoney, and P. Hospodka

The KEWI Research Group
College of Information Science and Technology
University Nebraska at Omaha
Omaha, Nebraska 68182

E-mail: wsousan@mail.unomaha.edu
Email: qzhu@mail.unomaha.edu
Email: rmnickell1@mail.unomaha.edu
Email: phospodka@mail.unomaha.edu
Email: wmahoney@mail.unomaha.edu

Abstract

The Internet offers a plethora of freely available information for possible use in Open Source Intelligence (OSINT) operations. However, along with this information come challenges in finding relevant information and overcoming information overload. This paper presents the results of an ongoing research in a Tailored Information Delivery Services (TIDS) system that aids users in retrieving relevant information through various open intelligence sources. The TIDS provides a semantics-based query constructor that operates in a "What You Get is What You Need (WYGIWYN^{TM})" fashion and builds ontology based information tagging, theme extractor, and contextual model.

Keywords: Information overload, information retrieval, open source intelligence, ontology, semantic annotation, ontological indexing, contextual query.

Introduction

Open Source Intelligence (OSINT) is the process of locating, harvesting, and analyzing information from publicly available sources such as the news media, public records, academic publications, and other data sources for intelligence collection purposes (Madill, 2005). The Internet offers a vast amount of data existing at international news sites, government reports, databases, and geospatial data that may be helpful in answering specific questions for military operations. Even though OSINT may not be the same as those obtained by classical methods such as covert operations, it still has the potential to provide useful information about emerging technologies, potential adversaries, operational environmental conditions, and supply information on operations that presently do not have classified intelligence. OSINT is defined by NATO as a strategic component in 21st century Information Operations (NATO, 2001) and envisions using OSINT activities for joint coalition operations. In addition to military operations, OSINT is useful to other types of organizations such as those that focus on humanitarian purposes by assisting with disaster relief operations. Examples of web sources that can provide open source information include the U.S. Foreign Broadcast Information Service (FBIS) which publishes news reports translated from various foreign-based media sources, and the British Broadcasting Corporation (BBC). Internet blogs owned by persons of interest also provide information that may contain possible intelligence value.

ISSN 1445-3312 print/ISSN 1445-3347 online

Several examples of systems developed for acquiring information from open sources exist in literature such as the iMiner project (Memon *et al*., 2007) which extracts and collects terrorist information from authenticated Web sites. In addition, the retrieved information is used for data mining purposes and is further used with graph theory for the analysis of terrorist networks. Likewise, OSINT is scanned and evaluated for the presence of potential terrorism warnings (Carroll, 2005) using artificial neural networks. As another example, an event-ontology is used by Palmer, (2005) to locate Actor-Action-Target event triplets for extracting event data from OSINT sources, where semantic comparators are used for measuring the semantic distance between triplets for event ranking.

The process of harvesting relevant data of a given event for evaluation in Requests for Information (RFI) tasks by intelligence analysts is a core component of OSINT data collection tasks. However, the overwhelming amount of information on the web continues to expand at an exponential rate, creating challenges for locating and retrieving information relevant to intelligence operations. As a result, these challenges establish the motivation for many researches in the development of intelligent systems for retrieving, filtering, and delivering user-tailored information from open sources.

In this paper, we discuss the results of ongoing research in the development of a Tailored Information Delivery Services (TIDS) system that is an extension of previous work presented at the International Conference on Information Warfare ICIW'08. The research can also be traced back to the Semantics Based Intelligent Web Service (SBIWS) project on semantic based information retrieval (Sousan *et al*., 2007). The nucleus of TIDS centers on a formal ontology that is used for query specification amongst users who share the need for the same information domain. In addition, retrieved text documents are converted to semantic net representations through theme extraction and tagged with concepts from the ontology that allow for them to be semantically indexed. The domain ontology can be extended and refined in an incremental learning process and is implemented in a standardized format that supports interoperability with external systems. Furthermore, users can create profiles that describe their individual information needs using concepts from the formal ontology in a context-sensitive format. As a result, relevant information is delivered that matches user’s profiles in a WYGIWYN^{TM} fashion.

The paper discusses ontology based information systems together with the system architecture and component description of the TIDS system. Subsequently, the operational system functionalities of TIDS are presented. Finally, current progress and challenges in the development of TIDS are discussed

Ontology Based Information Systems

An ontology is a model that describes concepts and their corresponding components such as the attributes, relationships between concepts, and may also contain rules that apply to the various components. As described by Gruber “A ontology is a formal explicit specification of a shared conceptualization” (Gruber, 1993). A core component of the semantic Web (Berners-Lee *et al*., 2001) is the use of ontologies for sharing information that is also machine interpretable. Ontologies provide a means for clearly describing concepts and their semantics unambiguously in a form of a knowledge model. It is, however, difficult to create and maintain ontologies, and much research has been spent in finding methods of building ontologies using automated, semi-automated, and manual methods of creation (Gruninger *et al*., 2002).

The use of ontologies in information systems provides several benefits. First and foremost, the knowledge needed and acquired in Ontology Based Information Systems (OBIS) can be stored in a standardized format that clearly and unambiguously describes the knowledge in a formal model. This model can be shared amongst other computer systems thus providing a means of knowledge sharing. In addition, software agents can interpret this knowledge and act upon it to assist humans in their knowledge operations. Second, ontologies may be hierarchical and thus provide a taxonomy of concepts that allows for the indexing of information. Thus, in information retrieval systems, the information retrieved can be indexed based on the ontology. Furthermore, the hierarchical structure allows the user to specify the information desired at different levels of granularity for a given concept.

Besides their retrieval characteristics, ontologies provide a means for data fusion by supplying synonyms or concepts defined using various descriptions. For example, the concept of “Fighter Aircraft” could be found in text under various descriptions as “F22-Raptor”, “F-16”, “Su-47 Berkut”, “JF-17”, “Chendgdu J-10” and other variations that all identify the same concept of a Fighter Aircraft. As a result, data from different sources can be collected by identifying the same concept. This feature is also useful by providing a means of semantically annotating ontological concepts within unstructured text. This allows the collection and indexing of unstructured text based on identified concepts. Thus users can clearly specify the concepts they desire in terms of ontological concepts, instead of keywords, and let the ontology determine which text strings correspond to a given concept. Furthermore, users can also retrieve information based on a set of concepts and their relationships, and thus specify a higher level of information needs in a contextual manner. For example, a user may need to retrieve information based on Bird Flu deaths of humans near lakes. This information would be difficult to express and retrieve accurately using simple keywords. However, using ontological concepts and relationships, the user could specify a contextual query based on a structure of concepts such as “Human Deaths From Avian Flu Near Lake” and the interest of the user recorded in his/her personal profile or previous search records. An onotology based system would be able to search a corpus of unstructured text looking for this described situation via contextual queries.

OBIS have been the topic of research in several different domains. For example, KIM (Kiryako *et al*., 2005) is used in retrieving names of people, organizations, locations, and other real-world entities. KIM uses ontology references for automated semantic annotation and supports semantic indexing and retrieval. Another example is AddMiner (Garcia *et al*., 2006), which performs text mining of offshore petroleum platform accident reports, using ontologies to create instances of the defined ontology. More recent examples of OBIS are the Metadata Extraction and Tagging Service (METS) system (Lee, 2007). METS produces large amounts of semantic information by converting heterogenous document formats into ontologies and annotating the corresponding metadata. In addition, ontologies are used for detecting possible threats by analyzing blogs in VIStology’s IBlogs project (Ulincy, 2007).

System Architecture

The TIDS is developed as a Web-based java application that runs on top of various open source software packages. To reduce development time and promote interoperability, TIDS uses various Commercial off-the-shelf (COTS) modules such as the Google Web Toolkit for the user interface, OWL (Dean M., 2002) ontology description language for the ontology implementation, and others. In addition, TIDS is constructed of several modules that run in a pipeline fashion to support the retrieval and delivery of information in accordance with users’ interests or designation. These components and their workflow are illustrated in figure 1.

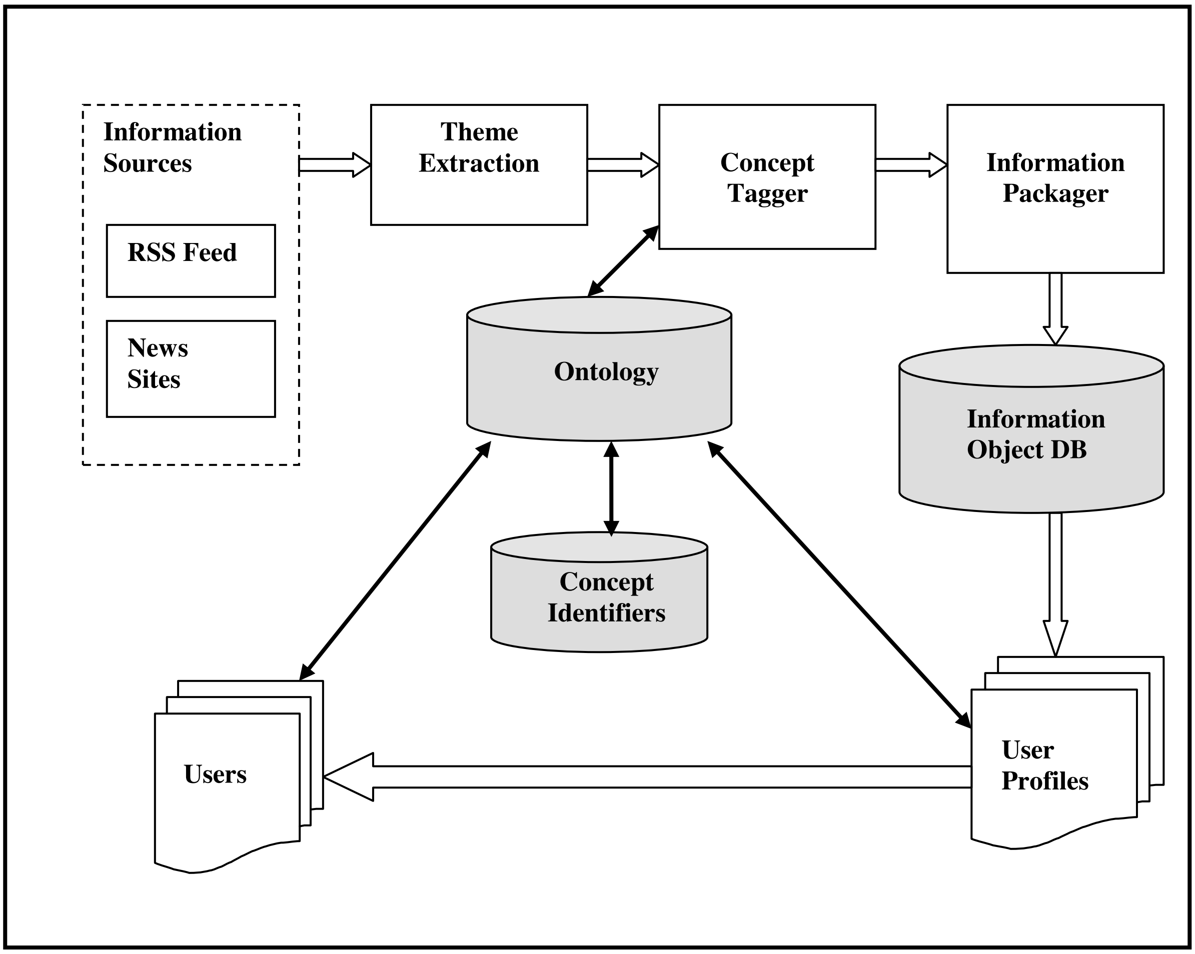

Figure 1: System Architecture

The following list describes the components and their responsibilities within the TIDS system architecture:

- *Information Sources*: Web sites that provide news from various world-wide events. For the initial prototype, sites such as the Google news archive and the BBC are used to provide test news reports in the form of unstructured text.

- *Theme Extractor:* This component is used to identify the theme or event domain of selected text-based news reports. This is done by extracting words and phrases which are later analyzed in order to determine whether they identify a concept defined within the ontology.

- *Ontology*: Repository of concepts and their relationships used by all to describe needed information. The ontology is also used to unify data across the system, and users can incrementally build and refine the ontology and add concept identifiers. Users select concepts from the ontology to create their specific profiles.

- *Concept Identifiers*: Consists of text patterns used for identifying ontological concepts.

- *Concept Tagger:* Used for identifying and tagging concepts from the ontology that is found within text.

- *Information Packager*: Used to store the information objects for distribution based on user's profiles. In addition, a semantic net representation is created to describe or summarize in ontological terms and relationships the overall context of the report. Thus a contextual model is created from the text of the news report.

- *User Profiles*: User configurations consisting of ontological concepts for use in specifying search criteria. These profiles are created through the selection of concepts and relationships from the ontology and are then registered with the system to indicate the type of information to be retrieved and delivered in terms of the user's interest.

System Operations

Ontology building

In TIDS, several military ontologies such as SUMO (Niles & Pease, 2001) and METS (Lee, 2007) and WordNet (Fellbaum, 1998) were analyzed in order to create a basic ontology consisting of various concepts such as transportation, weapons, personnel, military operations, and others. These initial concepts and terms are stored in a text file and can be converted into a corresponding ontology implemented in the Web Ontology Language (OWL) (Dean, 2002) which also provides for interoperability with other ontology-based systems. This ontology forms the nucleus of the system that provides a library of concepts that users can select from in order to create their profiles. These profiles, in consequence, form a type of user-defined mini-ontology which describes information needs in terms of ontological concepts in a contextual fashion. In order to prevent ambiguity and to provide for "unification" amongst users, the users are restricted to specifying their desired information needs from the ontology which can also be viewed as a metadata repository. Both the ontology and user's profiles can be built, extended, and refined in an incremental fashion similar to the author's work in SBIWS (Sousan *et al.*, 2007). Additional concepts, relationships, and identifiers can be inserted into the ontology and profiles as they are discovered or selected. As a result, a knowledgebase is formed in a collective intelligence manner in addition to the users' interaction in building their individual profiles. Collaborative processes in building ontologies have been found to be useful in other systems (Holsapple *et al.*, 2002). Presently the addition of concepts, relationships, and identifiers are manually added by users. However, future work will include an interactive process with the user that semi-automatically detects and add concepts. This may also include using WordNet (Fellbaum, 1998) for additional concepts and their corresponding synonyms, hypernyms, and hyponyms.

The concepts used in our initial ontology include those within the military domain, consisting of equipment, personnel, organizations and various relations that are arranged in a taxonomic structure. Figure 2 depicts a portion of our initial test ontology.

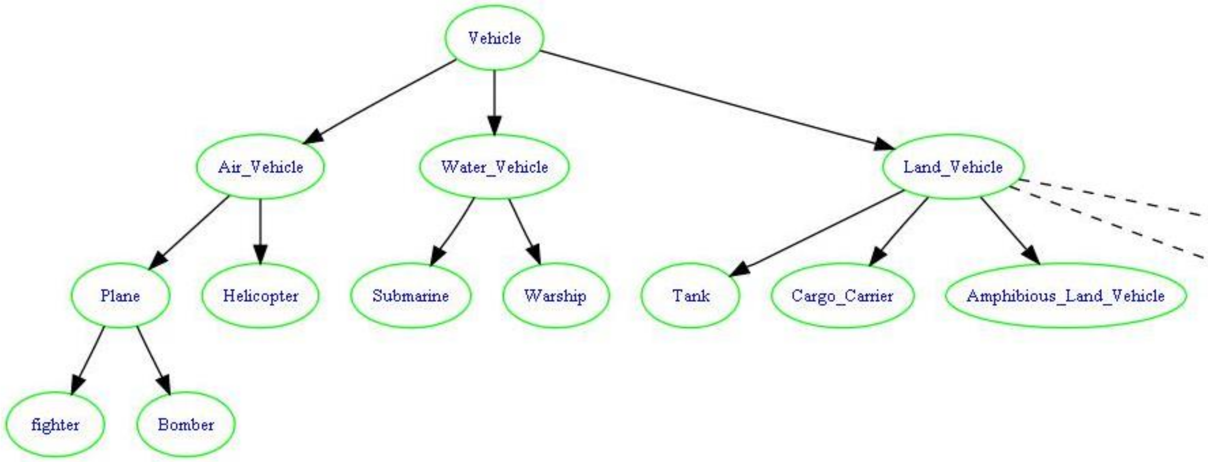

Figure 2: Section of the initial test ontology

Concept identifiers

For each concept defined in the ontology, there exists a corresponding identifier that consists of one or more words or phrases that are used to represent the concept. In addition, there may be a list of concept identifiers, such as synonyms or aliases, for the same concept that provides a means for fusing data from various information sources. These concept identifiers are sought for during the concept tagging process and are used to annotate sections of text with their corresponding concepts. For example, the concept identifiers "commando", "trooper", or "ranger" are all used to identify the concept of a soldier. Concept identifiers may be added as they are discovered during the text analysis process.

Contextual model

A contextual model is created based on the context of the news reports consisting of groups of concepts and relationships between them and stored in a semantic net or mini-ontology similar to the work in (Leskovec *et al*., 2005). This is done by syntactically analyzing the text and extracting parts of sentence sequences using nouns, verbs and adjectives for discovering richer relationships such as "Terrorist bombs airport" using similar techniques as described in (Palmer, 2005). These mini-ontologies are useful for describing such concepts, for example, as a military situation consisting of military aircraft attacks on small cities located on a country's border near water. These contextual models can be semantically compared to user's queries which can also be defined contextually. Figure 3 depicts the resulting semantic net representation from TIDS on a news report describing an intercontinental ballistic missile test by Russia.

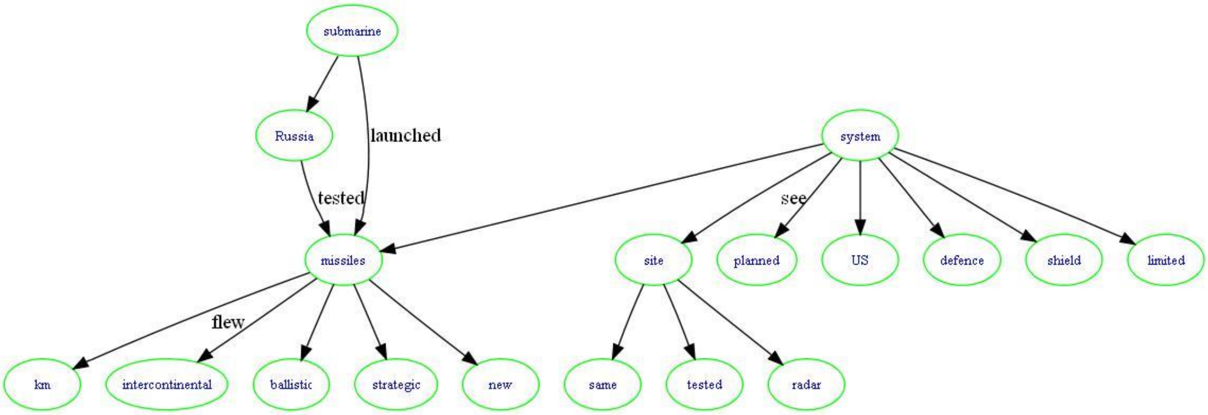

Figure 3: Semantic Net created from Text

Data harvesting

In TIDS, software agents are configured to continuously monitor web-sites whose content may contain information that semantically matches data described within the common ontology. As the pages are processed, they are scanned for the existence of concept identifiers by the theme extraction process to determine the overall theme of the text. If the theme is determined to be semantically similar to the common ontology, the text is annotated with its corresponding concepts and contextual model, thus adding descriptive metadata which describes the information in semantic terms. This process breaks down the text into a group of information units. The original text is retained along with the added annotations. As a result, this information can be analyzed later by querying it through ontological concepts.

For test purposes, we have manually selected a few news archive websites which provide relatively stable content collections. News sites such as Google news archive contains vast amounts of news articles that span several decades and provides test data for the evaluation of our system. Similarly, the BBC website and other various non-classified sites offer additional test data. To simplify the acquisition of test data, the text reports are extracted from the web sites and are placed into individual text files.

Tailored information delivery

Ontological concepts are used in each user's profile to describe their particular information needs, and can be viewed as a user-specified ontology for the user's query. The profiles are registered with the system to indicate what types of information are to be delivered to which users. In addition, the taxonomic level of the concept provides a means for the user to indicate the level of specificity that is needed. As the level of abstraction increases, the amount of returned information retrieved may increase as well, due to the larger number of sub-concepts in the concept hierarchy. In contrast, the more specific the concept is, the fewer the number of results will be returned, as there are fewer concepts within the concept hierarchy.

Discussion

The development of the TIDS system is presently in the early phases, and the focus is on the implementation and refining of components in support of collecting information containing possible intelligence value. To further understand the characteristics needed for the system and the initial concepts required, similar systems are being studied as well as upper-level and domain specific ontologies describing military concepts. Various military and terrorist events described in news articles, along with their context, are being tested in order to determine what type of concepts and relationships can be extracted from the articles. Similarly, different ontological structures are being evaluated to determine how best to structure them to model OSINT information. Building on the work of similar systems reviewed earlier, we are researching new methods of theme extraction, ontology modelling, semantic annotation, information packaging and their integration into the TIDS system.

As with other OBIS, our system faces several challenges that provide the opportunity for researching improved methods and novel solutions such as semantic annotation, semantic-based queries, ontology design and enhancement, and knowledge representation. To date the following issues have been identified and need further consideration:

- The structure of the current ontology is based on taxonomy of concepts, also known as a light-weight ontology, and is comprised of military concepts. However, more sophisticated ontology structures with relationships other than IS-A will offer greater knowledge modelling.

- News reports may be duplicated among various information providers, which need to be considered in order to avoid multiple copies of the same data. In addition, reports may contain varying levels of detail that could be combined into a single piece of information in order to get the complete picture.

- An inherit problem with collecting information from open sources is with its accuracy or trustworthiness. Problems arise with determining the level of accuracy as well as if the information was published for purposes of deception.

- Due to the multiple meanings of words, there exist challenges with word sense disambiguation. For instance the word "turkey" might indicate a bird or a country. We hope to reduce the disambiguation problems through the use of the theme extractor component.

- Problems may arise resulting from the collective process of users modifying the metadata repository. Users may have different points of view on the ontology structure, granularity, and categories.

- Initially the existence of user selected concepts is used to determine if information should be delivered to a given user. However, we are working on the ability to also specify relationships between concepts and thus specify a query on a contextual basis. This is the motivation behind the contextual queries.

Presently our system is under development and contains a small sub set of the desired features. More research needs to be completed so that the system can be of real value to OSINT operations. The following list outlines future capabilities:

- Ontology operations are limited to only adding concepts and concept identifiers. It may be beneficial to add support for pruning and/or re-arranging large sections of the ontology. Furthermore, users may want to visually review sections of the ontology for verification.

- Presently the system can not determine wether the same event is described by different reports. Therefore event tracking may be added to future versions that support an incremental process of collecting different attributes from the same event.

- To improve the usability of the system, the ontology building process will be converted to a semi-automated process in which the user interacts with the system to build and refine the ontology. Furthermore, the ontology can be extended with the analysis of delivered documents for additional extraction concepts and relationships.

Conclusion

In this paper we have described our ongoing research and challenges in the development of an ontological based information retrieval system, TIDS, using the OSINT domain for its environment of implementation and evaluation. Motivation for our research in TIDS is driven by the problems with current information retrieval systems in their inability to deliver relevant information and avoiding information overload. We reviewed the overall vision of our system including components that are required and not yet constructed as well as those components currently under development. The system serves as a platform for further research into the areas of ontology creation and refinement, contextual queries, semantic annotation, phrase and theme extraction, contextual modelling, and personalized information retrieval. As such, TIDS can be used in the front-end of the OSINT gathering process to collect raw data for intelligence analysts for further evaluation.

References

Berners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantic Web: A new form of Web Content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, 284(5): 28-37.

Carroll, J. M. (2005) OSINT Analysis using Adaptive Resonance Theory for Counterterrorism Warnings, International Conference on *Artificial Intelligence and Applications (AIA 2005*), Feb 14 - 16, Innsbruck, Austria, pp. 756-760.

Dean M., Connolly D., van Harmelen, F., Hendler J., Horrocks I., McGuinness D., Patel-Schneider P., and Stein L.A. (2002), *Web Ontology Language (OWL) Reference Version 1.0.* W3C Working Draft 12 Nov. 2002, URL: http://www.w3.org/TR/2002/WD-owl-ref-20021112/ [Accessed: 23rd August, 2008].

Fellbaum, C. (ed.) (1998) *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Garcia, A., Ferraz, I., and Pinto, F., (2006) The Role of Domain Ontology in Text Mining Applications: The ADDMiner Project, *Sixth IEEE International Conference on Data Mining – Workshops (ICDMW'06),* Dec 18, Hong Kong, China, pp. 34-38.

Gruber, T. (1993) A translation approach to portable ontologies, *Knowledge Acquisition,* **5**(2): 199-299.

Gruninger, M., and Lee, J. (2002) Ontology Applications and Design, *Communications of the ACM,* **45**(2):39-41.

Holsapple, C. W., and Joshi, K. D. (2002) A collaborative approach to ontology design, *Communications of the ACM* **45**(2): 42-47.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., and Ognyanoff, D. (2004) Semantic annotation, indexing, and retrieval, *Journal of Web Semantics,* **2**(1): 49-79.

Leskovec, J., Milic-Frayling, N., and Grobelnik, M. (2005) Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts, *National Conference on Artificial Intelligence (AAAI),* July 9-13, Pittsburgh, Pennsylvania, pp. 1069-1074.

Lee, R. (2007) The Use of Ontologies to Support Intelligence Analysis, *Proceedings of the Second International Ontology for the Intelligence Community Conference (OIC-2007)* November 28-29, Columbia, MD.

Madill, D. L. (2005) Producing Intelligence From Open Source*s, Military Intelligence Professional Bulletin*, **31**(2): 19-26.

Memon, N., Hicks, D. L., and Larsen, H. L. (2007) Harvesting Terrorists Information from Web, *Proceedings of the 11th international Conference information Visualization,* July 4-6, 2007, IEEE Computer Society, Washington, DC, pp. 664-671.

Niles, I., and Pease, A. (2001) Towards a Standard Upper Ontology, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19.

North Atlantic Treaty Organization (2001) *NATO Open Source Intelligence Handbook,* URL: http://www.au.af.mil/au/awc/awcgate/nato/osint_hdbk.pdf [Accessed: 23rd August, 2008].

Palmer, J. (2005) Textually retrieved event analysis toolset, *Military Communications Conference, MILCOM 2005*, **3**(Oct): 1679–1685.

Sousan, W.L., Payne, M., Nickell, R., and Zhu, Q. (2007) MetaData (Ontology) Incremental Building and Refinement Agents, *Proceedings of IEEE Knowledge Intensive Multiagent Systems (KIMAS '07*), April 29 – May 3, Waltham, Massachusetts, pp. 127-132.

Ulicny B., Matheus C., Kokar M., and Baclawski K. (2007) Uses of Ontologies in Open-Source Blog Mining, *Proceedings of the Second International Ontology for the Intelligence Community Conference (OIC-2007),* November 28-29, Columbia, MD.