WHITE PAPER

Semantic Enrichment and Fusion Of Multi-Intelligence Data

By Dr. Richard D. Hull, Don Jenkins, Alan McCutchen

Modus Operandi, Inc.
709 S. Harbor City Blvd., Suite 400
Melbourne, FL 32901
www.modusoperandi.com

Copyright
Copyright Information
Wave-EF White Paper
Copyright © 2006-2009 Modus Operandi, Inc.

Disclaimer
The sample data included in this guide is fictitious.

Trademark or Service Marks
Wave is a registered trademark of Modus Operandi, Incl. All trademarks are the property of their respective companies.

Table of Contents

1 Abstract

The challenges of predictive battlespace awareness and transformation of TCPED to TPPU processes in a net-centric environment are numerous and complex. One of these challenges is how to post the information with the right metadata so that it can be effectively discovered and used in an ad hoc manner. We have been working on the development of a semantic enrichment capability that provides concept and relationship extraction and automatic metadata tagging of multi-INT sensor data. Specifically, this process maps COMINT, IMINT and HUMINT data to concepts and relationships specified within a semantic model (ontology). We are using semantic enrichment for development of data fusion services to support multiple Department of Defense programs. This paper presents an example of using the semantic enrichment architecture for concept and relationship extraction from USMTF HUMINT reports and COMINT data. The process of semantic enrichment adds semantic metadata tags to the original data enabling advanced correlation and fusion. A geospatial user interface leverages the semantically-enriched data to provide powerful search, correlation, and fusion capabilities.

2 Introduction

Net-centricity transforms the traditional intelligence analysis process of Task, Collect, Process, Exploit and Disseminate (TCPED) to a Task, Post, Process and Use (TPPU) process. The fundamental difference between the two processes is that TPPU posts information before it is processed, making it available for other ad hoc purposes. To realize the value of this strategy, ad hoc users must be able to find relevant information from across the Department of Defense (DoD) Enterprise. All Net-centric systems have some mechanism for finding relevant information, typically content and metadata discovery services that require accurate and complete metadata tags to describe the content of information items they maintain.

We are developing a framework for semantically enriching sensor data and intelligence reports that can improve a net-centric system's discovery services by generating semantic tags for the information automatically, thereby helping users and applications find meaningful, situation-specific information more easily. The semantic tags, defined within one or more ontologies, represent the important concepts and relationships users need to fulfill their missions. This framework, called the Wave Exploitation Framework or Wave-EF, is built upon a publish and subscribe messaging service, an open source unstructured information management system, and components for transforming the raw information into XML and generating semantic tags.

Wave-EF is designed to function within a service-oriented architecture (SOA), with rapid integration into Enterprise Service Bus (ESB), Metadata Registries and other components of a SOA ecosystem. It is both a consumer and producer of enterprise and web services. The use of a publish/subscribe mechanism means that the Wave-EF services do not need be collocated, an important element of the net-centric and TPPU philosophies.

Wave-EF was used recently to develop a pipeline for semantically enriching United States Message Text Format (USMTF)[1] reports with metadata tags relating to vehicle thefts. The use case we were addressing was: "Intelligence analysis has identified a number of potential indicators of a future detonation of a vehicle borne improvised explosive device (VBIED), one of which is the theft of an appropriate vehicle for use in the VBIED attack. How can we identify vehicle theft events described within intelligence reports with high precision?" This manuscript describes our methodology and results towards recognizing descriptions of vehicle theft events within USMTF reports and creating the appropriate semantic tags, i.e., tags representing the concepts and relationships of each 'true' event. This capability, once fully generalized, will provide automated tagging of military information, thereby accomplishing two critical objectives simultaneously: (1) accelerating the pace with which information can be made accessible by net-centric discovery services; and (2) enhancing the meaning of these messages for more accurate filtering, fusion and situation-specific analysis.

Discovery services

Discovery services in a net-centric environment provide visibility, accessibility and understandability of data and service assets across the DoD Enterprise. According to the DDMS Specification [2], "The Department of Defense Discovery Metadata Specification (DDMS) defines discovery metadata elements for resources posted to community and organizational shared spaces. 'Discovery' is the ability to locate data assets through a consistent and flexible search. The DDMS specifies a set of information fields that are to be used to describe any data or service asset that is made known to the Enterprise, and it serves as a reference for developers, architects, and engineers by laying a foundation for Discovery Services. The DDMS will be employed consistently across the Department's disciplines, domains and data formats."

The DoD Net-Centric Data Strategy (NCDS) and DoD Directive Number 8320.2 require data sharing across the DoD, including the creation of new information resources to describe the available assets:

> [POLICY] 4.2. Data assets shall be made visible by creating and associating metadata ("tagging"), including discovery metadata, for each asset. Discovery metadata shall conform to the Department of Defense Discovery Metadata Specification (DDMS)[3].

The recommended best practice for describing an asset is to associate it with one or more DDMS Subject Content Category metadata entries from a controlled vocabulary. Communities of Interest (COIs) are responsible for defining the controlled vocabularies, taxonomies or classification schemes for their respective domains and ultimately for overseeing the use of those schemes during metadata tagging. Unfortunately, the manual tagging of data and service assets is tedious, error prone and resource-intensive. Our work is aimed at automating these tasks.

Semantics

Decades of research in knowledge representation and machine reasoning has led to recent efforts in the creation of a number of robust technologies for implementing the 'Semantic Web'.[4] In particular, the DARPA-sponsored efforts including the DARPA Agent Markup Language (DAML)[5] and the Ontology Inference Layer (OIL)[6], collectively known as DAML+OIL, were used along with the Resource Description Framework (RDF)[7] to create the World Wide Web Consortium's (W3C) Web Ontology Language or OWL[8]. OWL provides a standards-based language for defining ontologies of concepts and relationships, definition of properties of concepts and relationships, set theoretics (e.g., union, intersection) and equivalence constructs between concepts and individuals (concept or class instances).

OWL as a representation language for COI taxonomies and controlled vocabularies has been used in the DoD's Metadata Registry [9]. Therefore automatic production of OWL metadata tags would support the COI efforts across the DoD. One other large effort underway to produce OWL metadata tags is the Automated Metadata Population Service (AMPS).[10] Our work differs from AMPS in that we are focusing on the generation of OWL-based subject content tags instead of keywords.

Semantic enrichment

Semantic enrichment is the process of adding or associating semantic tags – usually concepts, relationships, events and properties described in an ontology – to augment unstructured data items. The semantic tags are persisted and used for the search and retrieval of the original data items as well as for machine reasoning over the tags themselves. We are currently developing semantic enrichment capabilities to support multi-intelligence (multi-INT) data fusion for DoD C4ISR applications by tagging raw and exploited HUMINT, IMINT and COMINT with Joint Command Control and Consultation Information Exchange Data Model (JC3IEDM)[11] and Universal Core (UCore)[12] concepts and relationships.

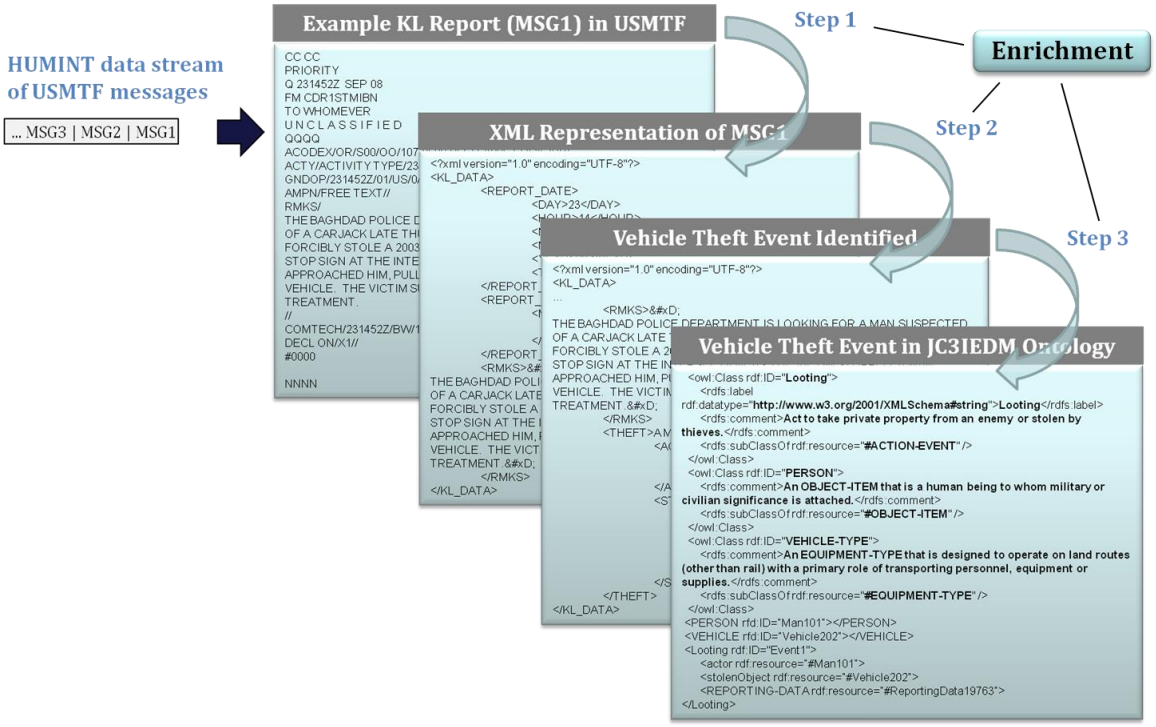

Fig. 1. Example of HUMINT/COMINT data streams (USMTF messages) are first transformed into XML (Step 1). Vehicle theft events are recognized and XML tags representing them are added (Step 2). These elements are then transformed into OWL ontology constructs (Step 3).

An example of the process of semantic enrichment (shown in Figure 1) begins with the transformation of raw source data into an intermediary eXtensible Markup Language (XML)[13] format. This step identifies the major structural features of the source data, including header, body and footer features. We have developed tools that allow us to quickly develop a data extraction language (DEL) for the incoming source. A source can be textual or binary and structured, semi-structured or unstructured.

During the second step of the transformation process, each structural feature is scanned so that domain features can be identified and annotated. If the source information is a USMTF message, a handler specific to each structural field of the message is invoked. The 'location' field of a USMTF message can be converted from Military Grid Reference System (MGRS)[14] coordinates to geodetic latitude and longitude coordinates during transformation into XML. The 'remarks' field is forwarded to an entity extraction system to identify entities (concepts) of interest. This flexible approach is also quite powerful and has allowed us to leverage both proprietary and third-party (open source, COTS, GOTS) components for efficient transformation of source data.

The second step produces XML annotations for salient features identified in the source data; it produces no annotations if the message does not describe a theft event or a related concept or relationship. The third step is to map the extracted features to the concepts and relationships defined in the target ontology or ontologies. For example, the application we will describe in Section 0 involves the identification of vehicle theft events in USMTF messages. The XML representation for a vehicle theft event is a <VEHICLE-THEFT> element. If the target ontology is JC3IEDM-based, then there are two potential concepts we could map to: a subclass of ACTION-EVENT called 'Looting' or a subclass of ACTION-EVENT called 'Robbery'[1]. We have chosen to map to the 'Robbery' subclass when it's clear that the theft

[1] JC3IEDM is not an ontology but a data model. There have been efforts to construct an OWL ontology from JC3IEDM using XSLT [20]. However, the Looting and Robbery ACTION-EVENTs in this JC3IEDM ontology do not capture the semantic cases necessary for

was directly from a person rather than an organization, group or location, e.g., because the verb 'carjack' or 'hijack' was used or if the car was stolen directly 'from' a person. All other instances map to the 'Looting' concept.

We have studied the problem of automating the mapping of schemas and ontologies and while it is difficult, there are alternative approaches that can provide useful solutions. Rahm and Bernstein surveyed the state of the art in automatic schema matching and found that there are a number of different approach types, some more suited than others for semantic query processing [15]. While this survey is directed more towards relational databases, many of the ideas have merit in the realm of ontology concept matching, in particular, the linguistic approaches. Another recent work is C-OWL, which uses a set of bridge rules to create a mapping between two ontologies [BOU2003]. The bridge rules "define semantic relations between concepts in different ontologies." This work builds upon research published by Borgida and Serafini regarding distributed description logics, which extends the reasoning available on ordinary schemas to the case of multiple schemas connected by arbitrary binary correspondences between individuals (i.e., bridge rules) [BOR2002].

For the purposes of this discussion, we will assume that a mapping between the XML elements and the target ontology has been created.

Wave-EF

The Wave Exploitation Framework was used as the semantic enrichment platform for the experiment described here. As shown in Figure 2, Wave-EF uses a publish and subscribe messaging system (currently the Java Message Service[18], but it can be easily swapped with other publish/subscribe systems such as the Data Distribution Service[19]) to route incoming multi-source intelligence data through a semantic enrichment pipeline to a number of applications and services.

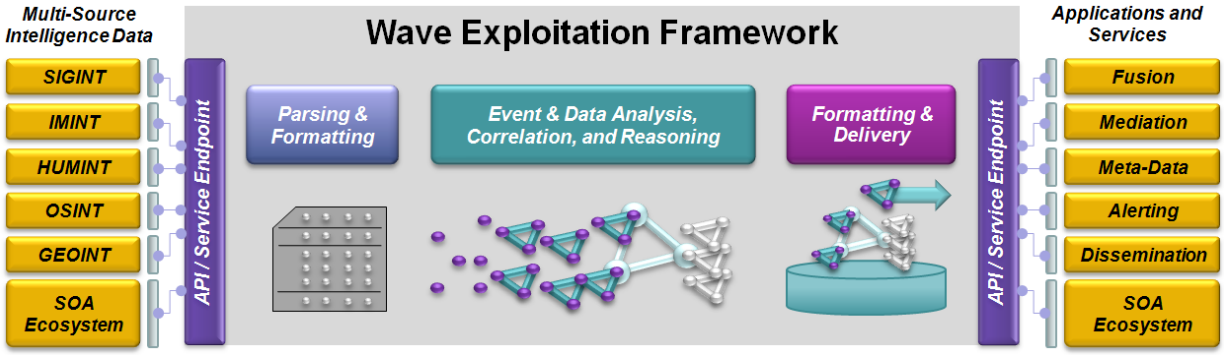

Fig. 2. Wave-EF semantically enriches multi-source intelligence data through direct feeds or from net-centric services for fusion, mediation, metadata tagging, alerting and dissemination within applications and services.

Processing is typically from left to right through the pipeline, but there is no requirement that all steps be executed and other orderings are possible. Formatting and parsing of the incoming data, i.e., translation into XML, is handled by two components, Wave Formatter and Wave Parser. Wave Formatter allows a knowledge engineer to define a DEL mapping to XML and later Wave Parser uses that specification to translate items as they are published. The concept and relationship extraction steps are performed within the Apache Unstructured Information Management Architecture (UIMA)[21], a Java-based, open source architecture for processing unstructured information. We are using UIMA to organize a number of concept, relation and event extractors, or as they are called in UIMA, annotators. These annotators can detect and extract concepts (or entities) including simple concept instances such as specific locations,

representing these events. We have added cases including perpetrator, theme, location and time to our own OWL representation of JC3IEDM.

dates, times, email addresses, as well as more complex items such as JC3IEDM concepts or vehicle theft events. We have built our own powerful annotators using a combination of regular expression patterns and semantic variables. These extensions to UIMA are described in more detail in Section 0.

The results of the extraction process are then translated into OWL instances or RDF triples and persisted for further correlation and fusion, and published to the Enterprise for use by interested (subscribing) applications and services.

Identifying vehicle theft events

We have used Wave-EF to address a representative problem faced by analysts looking for very specific events and concepts within a large corpus of potentially similar yet unrelated data. The problem involves detection of indicator or precursor events to VBIED attacks within HUMINT reports. The goal of this effort was to accurately identify and extract descriptions of the thefts of vehicles in and around an Area Of Responsibility (AOR) that may precede a car or truck bomb attack on US military and civilian targets. Currently, this is being done manually by analysts reading all incoming reports of criminal activity or by using traditional search engine technology to search for reports that mention keywords like “steal”, “stole”, “stolen”, “car”, “truck”, “vehicle”, etc.

Even using search tools, the analysts had to review many irrelevant stories because their tools could not differentiate among semantically relevant and irrelevant stories. Moreover, important reports were missed because of incomplete keywords. Therefore, we identified this gap as an opportunity to use Wave-EF to semantically tag incoming reports with JC3IEDM representations of vehicle thefts in the reports that describe them. These automatically-generated tags are persisted in the Wave-EF knowledge base, but they could just as easily be used to populate metadata entries in a net-centric metadata catalog. While this paper describes the methodology and results for identifying vehicle theft events, our approach can be generalized to identify other JC3IEDM event types. In fact, we are currently defining extraction patterns for other event types and plan to have patterns for all 347 JC3IEDM enemy event types in the future.

3 Methodology

The semantic enrichment process used in this project involved the following steps: 1) identification and parsing of the various fields of the USMTF messages; 2) extraction of concepts and relationships found in those fields and the generation of the appropriate XML tags for them; and 3) translation of the XML tags into JC3IEDM OWL representations. Steps 1 and 3 are straightforward as they both involve manipulation of structured formats. Step 2 is more difficult because it requires analysis of unstructured text to extract concepts, relationships and, ultimately, events. Therefore we will focus the following discussion on the methodology for implementing Step 2.

Extraction patterns

Extraction of concepts, relationships and events requires a systematic description of the information to be extracted, i.e., a set of extraction patterns. To build these patterns, a knowledge engineer created a training set of 32 USMTF messages. These messages contained fictitious, unclassified HUMINT reports of vehicle thefts and similar criminal events. Manual analysis of the training set messages and similar news reports by the knowledge engineer resulted in a number of logical patterns describing vehicle theft events (similar to a phrase structure grammar). Examples of patterns for active and passive voice sentences using the past tense of the verb ‘steal’ include:

Active voice: <person | organization> [adverb] stole <vehicle> [from <location>] [adverb] (1)

Passive voice: <vehicle> was stolen [from <location>] [by <person | organization>] [adverb] (2)

Items in angle brackets (<>) are classes or kinds in our ontology: they represent any word or phrase denoting an instance of the class, e.g., a person (an individual's name or nouns such as 'man' and 'thief'), an organization, a location, etc. Items in square brackets ([ ]) are optional. Parts of speech, such as 'adverb', are used to match instances of that part of speech, e.g., 'allegedly', 'reportedly' or 'yesterday'. A vertical (|) represents a logical disjunction, e.g., < person | organization > matches instances of a person class or an organization class. Active voice pattern (1) matches sentences that describe a person or organization stealing a vehicle with an optional adverb and an optional prepositional phrase indicating from where the vehicle was stolen.

The two patterns above identify sentences describing vehicle theft events: "John Smith stole a truck," "A terrorist group allegedly stole the fuel truck from the Oil Ministry" and "A 2003 Chevrolet Suburban was stolen yesterday." Patterns for matching vehicle theft events using other tenses of the verb steal and other verbs were also constructed. All of these patterns form two *über* [2] patterns as the verbs of 'stealing' tend to the same verbal subcategorizations[3]. Therefore we define these two *über* patterns to be:

Active voice: <person | organization> [adverb] <steal> <vehicle> [from <location>] [adverb] (3)

Passive voice: <vehicle> was <steal> [from <location>] [by <person | organization>] [adverb] (4)

The new item <steal> represents the different verbs depicting the theft of an object, e.g., take, hijack, carjack, make off with, pilfer, etc. Other prepositional phrases not shown here but also part of the *über* patterns include prepositions for spatial relationships ('near', 'by', 'within') and prepositions for temporal relationships ('during', 'on', 'before', 'after'). The use of externally manageable dictionaries further enhances the power of the *über* patterns. For example, references to specific makes and models of vehicles (e.g. "Chevrolet Suburban") can be matched using a <vehicle> dictionary. As new makes and models emerge, only the external dictionary is changed, while the *über* patterns remain the same.

From each of these *über* patterns, a regular expression was created to recognize instances of the pattern within intelligence reports. According to Wikipedia, "In computing, regular expressions provide a concise and flexible means for identifying strings of text of interest, such as particular characters, words, or patterns of characters."[22]. The regular expressions are used to match the appropriate substrings (portions of the HUMINT sentences) and to extract the semantic cases (i.e., thematic roles) of the vehicle theft event. For example, considering the sentence "John Smith stole a truck", the semantic case representing the actor or perpetrator of the theft event is "John Smith" and the semantic case representing the theme or object stolen is "the truck". Identification of these semantic cases is crucial for subsequent analysis of the data.

Using these *über* patterns, the header and body sections of the original 32 training messages and 168 new test messages, describing vehicle theft and other events, were analyzed to detect and capture vehicle theft events. We calculated recall and precision metrics for accurate detection and extraction of vehicle thefts and compared them against the results from using a search engine with a complex Boolean query. The query we used was:

(stole OR stolen) AND (vehicle OR car OR truck OR suv OR van OR bus)

This is actually a more complex query than most analysts would create. Training new analysts in the complexities of Boolean queries takes additional time and effort. Moreover, this illustrates a weakness of keyword queries: one must know and be prepared to type all of the possible verbs and vehicle types that could be used in a message. Our approach, however, can leverage all of the subtypes (e.g., 'car' and 'truck') and instances of <vehicle> stored within the ontology or separate concept dictionaries.

The semantic enrichment results were displayed within a geospatial visualization tool. The tool also provides filtering by geographic region and date/time, used to further reduce the output. For example, the analyst can request *only those reports that indicate a vehicle theft in a specific region of the AOR during the last fifteen day*s. Locations and times of the thefts are extracted from the USMTF header information.

---

[2] The term *über* here is used to express a super pattern which subsumes all others patterns for a particular voice.

[3] Verbal subcategorizations describe the number and type of syntactic arguments that a verb co-occurs with. Two major subcategories of verbs are the transitive (verbs which take a direct object as in 'Peter ate an apple') and the intransitive (verbs which do not take a direct object as in 'Peter sleeps') verbs.

4 Data

Our experiment involved the use of 200 USMTF messages describing real events from in and around an AOR. An example of a USMTF message portraying a vehicle theft event is shown in Figure 3.

```
CC CC
PRIORITY
Q 231452Z SEP 08
FM CDR1STMIBN
TO WHOMEVER
U N C L A S S I F I E D
QQQQ
ACODEX/OR/S00/OO/107,0500Z/FCH,A00F,SRD:SXA//
ACTY/ACTIVITY TYPE/231452Z//
GNDOP/231452Z/01/US/0/-/0/MG:38SMB2673486211/ELP:0.3KM-0.2KM-340.2//
AMPN/FREE TEXT//
RMKS/
THE POLICE DEPARTMENT IS LOOKING FOR A MAN SUSPECTED OF A CARJACK LATE
THURSDAY NIGHT.  AT ABOUT 11:45 P.M., A MAN FORCIBLY STOLE A 2003
CHEVROLET SUBURBAN WHEN HE CAME TO A STOP SIGN AT THE INTERSECTION. AS THE
VICTIM STOPPED, A MAN APPROACHED HIM, PULLED HIM FROM HIS VEHICLE AND
STOLE THE VEHICLE.  THE VICTIM SUFFERED MINOR INJURIES BUT REFUSED
TREATMENT.//
COMTECH/231452Z/BW/145.00MHZ/MD:S/PSD:XX/TL:231452Z/CS:BG777//
DECL ON/X1//
#0000
NNNN
```

Fig. 3. An example of a USMTF message describing a vehicle theft.

We purposefully added messages containing 'red herring' events to the test set. Red herring events are those that mention keywords associated with vehicle theft events but do not describe an actual event. For example, a message containing the sentence, "A car wash was found using water stolen from broken city pipes," contains the keywords 'car' and 'stolen' but does not describe the theft of a car. Red herring messages are erroneously retrieved, however, by traditional search engines and can cause intelligence analysts to waste their time reviewing them. Our approach avoids this pitfall because the *über* patterns define contexts for the keywords that must be present to convey a true vehicle theft relationship.

5 Results

Of the USMTF 200 messages, 35 describe vehicle theft events. Precision, recall and F-measure statistics are shown in Table 1. Precision is the ratio of the number of relevant messages selected to the total number of messages selected. Recall is the ratio of number of relevant messages select to the total number of relevant messages. F-measure is the weighted harmonic mean of precision and recall and is used to combine both precision and recall into a single score. In all cases, larger numbers are better.

Table 1. Comparison of Wave-EF, search engine and random precision, recall and F-measure statistics.

	Msgs Selected	Relevant Msgs Selected	Precision	Recall	F-measure
Wave-EF	**23**	**23**	**1.00**	**.657**	**0.793**
Keyword Query	**67**	**29**	**.433**	**.829**	**0.569**
Random	**97**	**17**	**.175**	**.489**	**0.223**
Select All Msgs	**200**	**35**	**.175**	**1.00**	**0.241**

Wave-EF was the most precise method by a wide margin. The F-measure score for Wave-EF was significantly better than the other approaches. However, Wave-EF's recall number was below that of the keyword query and the selection of all messages. The decline in recall was due to 12 messages containing vehicle theft events which were not identified by Wave-EF. The sentences or clauses that should have led to a correct identification, but were missed by Wave-EF, are shown in Table 2 with the reason for the missed identification. Most of these missed events can be fixed with updates to the concept dictionaries or minor changes to the *über* patterns.

ID	Description	Reason
vt8	... when a vehicle and trailer were stolen ...	conjoined subject
vt21	The car was originally reported stolen from a nearby neighborhood.	complex adverbial
vt25	The car used in the bombing was stolen from a church there.	embedded relative clause
vt42	Last week, a gang stole a truck with four cows in it.	missed subject
vt47	... 32 suspects were captured after stealing two cars ...	long distance dependency
vt56	... noting that they stole the vehicle and detained him.	anaphora
vt58	The insurgent fighters had commandeered the vehicle from two soldiers	missed verb
vt72	A state-owned pick-up, which had been stolen from a state department, ...	missed subject
vt79	The appeals court also fined each of them for stealing a car ...	missed subject
vt82	Several police and army vehicles had been stolen	conjoined subject
vt100	The car was hijacked from its owner a day earlier.	missed verb
vt105	They are stealing everything: fire trucks, water tankers, ...	missed object

These results are displayed within a geospatial visualization tool using latitude/longitude pairs extracted from the message. The simple demonstration tool shown in Figure 4 illustrates the results of Wave-EF filtering. The map is dramatically less cluttered than it would be if the 200 original or even the 67 keywords results were displayed. The analyst can select any of the small blue vehicle icons to see the original message and the extracted vehicle theft event.

Fig. 4. These screenshots show the original 200 events (left) and the results after filtering the semantically enriched events for vehicle thefts (right). The resulting events are projected as icons onto a map of the AOR for analysts to examine. Selecting an event icon causes the message details and the semantic metadata to be displayed.

These results are very promising and show that Wave-EF can identify vehicle theft events with high precision. In addition, key attributes that greatly enhance discovery of relevant data are also extracted. These attributes include the time, location, perpetrator and stolen object attributes of these events. The extracted attributes can be used as semantic metadata tags associated with the messages and stored in a net-centric metadata catalog or to populate a RDF triple store for reasoning. For example, the JC3IEDM metadata tags generated for USMTF message described in Figure 3 are shown in Figure 5.

```
<owl:Class rdf:ID="Looting">
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Looting</rdfs:label>
    <rdfs:comment>Act to take private property from an enemy or stolen by thieves.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#ACTION-EVENT" />
 </owl:Class>
 <owl:Class rdf:ID="PERSON">
        <rdfs:comment>An OBJECT-ITEM that is a human being to whom military or civilian significance is attached.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#OBJECT-ITEM" />
 </owl:Class>
 <owl:Class rdf:ID="VEHICLE-TYPE">
     <rdfs:comment>An EQUIPMENT-TYPE that is designed to operate on land routes (other than rail) with a primary role of transporting personnel, equipment or supplies.</rdfs:comment>
    <rdfs:subClassOf rdf:resource="#EQUIPMENT-TYPE" />
 </owl:Class>
 <PERSON rfd:ID="Man101"></PERSON>
 <VEHICLE rfd:ID="Vehicle202"></VEHICLE>
 <Looting rdf:ID="Event1">
    <actor rdf:resource="#Man101">
    <objectStolen rdf:resource="#Vehicle202">
    <REPORTING-DATA rdf:resource="#ReportingData19763">
</Looting>
```

Fig. 5. Semantic metadata tags generated for a vehicle theft event.

The event extracted is an instance of a Looting event with the identifier 'Event1'. The actor (i.e., the agent performing the looting or the perpetrator) of the Looting event is 'Man101'. Man101 is an instance of PERSON class. The

identifier ‘Man101’ is automatically generated by concatenating the class name with an increasing integer. We use this scheme for readability, but any unique identifier could be used (e.g., GUID). The object stolen of the Looting event is ‘Vehicle202’. Vehicle202 is an instance of the VEHICLE class. Time and location information is stored in the REPORTING-DATA class instance with the identifier ‘ReportingData19763’. The REPORTING-DATA class contains information about the date, time, reliability, source, etc. Event locations are not shown here, but are captured as well.

6 Conclusions

We have shown how HUMINT reports can be semantically enriched with metadata tags from a military ontology (JC3IEDM) for identifying vehicle theft events. These events represent potential precursors to VBIED attacks on US and Allied forces in and around an AOR, and intelligence analysts need a way to retrieve them from the much larger sets of reports found in DoD repositories (e.g., the DCGS-A Brain). While our experiment’s combined training and test set was only 200 messages, the actual number of local police and US military messages is much larger. We are currently developing new methods for extracting events that can handle very large numbers of input messages, on the order of 1000’s to 10,000’s per day. With that volume of incoming information, the ability to identify and extract events with high precision is critical. Furthermore, because we intend to apply this approach to automatic metadata tagging of net-centric information, the tags we generate must be as correct as possible. They must also be transformed into a standard representation that allows them to be combined with additional information about the same events, through an iterative semantic enrichment process.

7 Acknowledgements

The authors would like to acknowledge the contributions of the Wave-EF development team who have helped both in the development of Wave-EF and in the accuracy of this manuscript. They are Dan Nieten, Rob Asfar, Mike Gill and Javad Maharramzade. Teresa Nieten developed the software for projecting vehicle theft events onto a geospatial display and in defining geospatial rules for event filtering.

8 References

[1] DoD MIL-STD-6040 US Message Text Format (USMTF).

[2] Department of Defense Discovery Metadata Specification (DDMS), VERSION 2.0, July 16, 2008. http://metadata.dod.mil/mdr/irs/DDMS/.

[3] Department of Defense Directive Number 8320.2 (December 2, 2004), p. 2, directive certified current as of April 23, 2007.

[4] Feigenbaum, L., Herman, I., Hongsermeier. T., Neumann. E, and Stephens, S., "The Semantic Web in Action," Scientific American, vol. 297, 90-97, (2007).

[5] Hendler, J. and McGuinness, D. The DARPA Agent Markup Language. IEEE Intelligent Systems, 15(6), 67-73 (2000).

[6] Fensel, D., et al. "OIL: An Ontology Infrastructure for the Semantic Web," IEEE Intelligent Systems, 16(2), 38-45, (2001).

[7] Resource Description Framework (RDF). W3C, http://www.w3.org/RDF/, (2001).

[8] OWL, Web Ontology Language. W3C, http://www.w3.org/TR/owl-features/, (2004).

[9] DoD Metadata Registry Status, Dr. Glenda Hayes, NCES Data Advisor, May 17, 2006.

[10] Automated Metadata Population Service (AMPS), Spiral 1 Workshop, M. Uhart, V. Dobbs, October 30, 2008.

[11] Joint C3 Information Exchange Data Model v3.1b (JC3IEDM Main), Multilateral Interoperability Programme, Greding, Germany, December 2007.

[12] The Universal Data Core, Army Net-Centric Data Strategy (ANCDS), http://data.army.mil/datastrategy_universal_core.html, (2009).

[13] eXtensible Markup Language, XML. W3C, http://www.w3.org/XML/, (2008).

[14] DMA Technical Manual 8358.1, Chapter 3. Datums, Ellipsoids, Grids, and Grid Reference Systems, (2006).

[15] Rahm, E. and Bernstein, P.A. A survey of approaches to automatic schema matching .The VLDB Journal 10: 334–350, (2001).

[16] Bouquet, P., et al. "C-OWL: Contextualizing Ontologies." Second International Semantic Web Conference (ISWC-2003), LNCS vol. 2870, 164-179, Springer Verlag, (2003).

[17] Borgida, A. and Serafini, L. "Distributed Description Logics: Directed Domain Correspondences in Federated Information Sources," In Proceedings of the International Conference on Cooperative Information Systems, (2002).

[18] Java Message Service. (2009, Jan 12). In *Wikipedia, the free encyclopedia*. Retrieved Jan 20, 2009, from Wikipedia, http://en.wikipedia.org/wiki/Java_Message_Service.

[19] Data Distribution Service for Real-Time Systems Specification, http://www.omg.org/docs/ptc/04-03-07.pdf.

[20] Matheus, C. J. and Ulicny, B. "On the Automated Generation of an OWL Ontology based on the Joint C3 Information Exchange Data Model." 12th ICCRTS, Newport, RI, June 19-21, (2007).

[21] Unstructured Information Management Architecture. http://incubator.apache.org/uima/index.html.

[22] Regular expression. (2009, Jan 19). In *Wikipedia, the free encyclopedia*. Retrieved Jan 22, 2009, from Wikipedia, http://en.wikipedia.org/wiki/Regular_expression.