Blending Concept Extraction, Semantic Wikis and Ontologies for Text Analysis: The Luminary System

Edward Swing, VSTI/SAS

Abstract—Automated text analysis encompasses a number of different techniques. However, applications frequently implement these techniques as single solutions, rather than creating an integrated environment that includes interaction and representation for the analyst. The Luminary system blends semantic reasoning, entity and concept extraction, entity co-referencing, and visualizations, all controlled by a governing domain ontology. It stores the results of its analysis in a semantic wiki, providing a familiar yet powerful user interface to analysts. At the same time, the wiki API enables the Luminary wiki to interact with automated processes. This innovative design enables analysts to modify the automated analysis process, review and correct inaccuracies in the results, collaborate with other analysts, perform sophisticated semantic queries, and publish analytical results.

Index Terms— Semantic Wikis, Text Analysis, Entity Extraction, Semantic Content Extraction, Document Categorization, Entity-Relation Modeling, Knowledge Representation.

INTRODUCTION

Over 80% of the world's information resides in unstructured text documents [3]. This vast volume of information presents many challenges to automated systems, which have wrestled with the challenges of unstructured text for years. The field of text analysis has generated many different algorithms, techniques, and concepts attempting to mine useful information from the morass of text. Entity extraction, content categorization, sentiment analysis and text clustering algorithms all attempt to provide a structure to the documents that an automated system can comprehend.

A number of systems and utilities offer entity extraction techniques. These systems can detect persons, locations, organizations, and other entities within documents. Most entity extraction engines also provide entity types, indicating whether a particular entity is a person, organization, etc. However, despite these capabilities, even mature entity extraction engines have significant error rates.

Since both entity and concept extraction still struggle with precision and accuracy, any automated extraction process will require some amount of human review and validation. Developers have created a variety of custom applications for text analysis and review, each incorporating different capabilities and techniques.

Recent advances in semantic web technology offer new approaches for text analysis and knowledge representation on the web. Wikis, such as Wikipedia, offer a simple, consensual knowledge base that millions of people use daily. Wikipedia has become a recognized information source worldwide. Many automated systems use the contents of Wikipedia for text content, references, or even text categorization challenges (Gabrilovich and Markovitch, [2]).

Other researchers, such as Guarino [4], have explored how to apply ontologies to knowledge engineering challenges. Exploring diverse topics such as epistemology and linguistic concepts, these works provide some formalism and theory to knowledge representation challenges. Researchers, such as Agichtein [1] and Kokar [6], have also investigated how to apply ontologies or similar semantic techniques to concept extraction. Naturally, some developer teams, such as the OntoWiki project [5], have developed wikis to support their ontological research.

The Luminary project blends multiple technical approaches for entity and concept extraction into a cohesive process, and integrates novel semantic algorithms for exploring information. It uses multiple entity extraction engines in parallel, and adds an Entity Verification step to validate the results of entity extraction. It focuses the concept extraction using a set of semantic lexical parsing rules. It uses a semantic wiki as a concept repository, thus offering editing, consistent knowledge representation, and analytical collaboration. Finally, it coordinates both the semantic extraction process and the contents of the semantic wiki through a governing OWL ontology.

1 THE EXTRACTION PROCESS

Vision Systems & Technology Inc., a subsidiary of SAS Inc., developed the Luminary prototype system. The Luminary system extracts information from news articles on public websites, RSS feeds, or documents in Word or PDF format. It runs through several automatic analysis steps, and pushes the results of the text analysis into a semantic wiki.

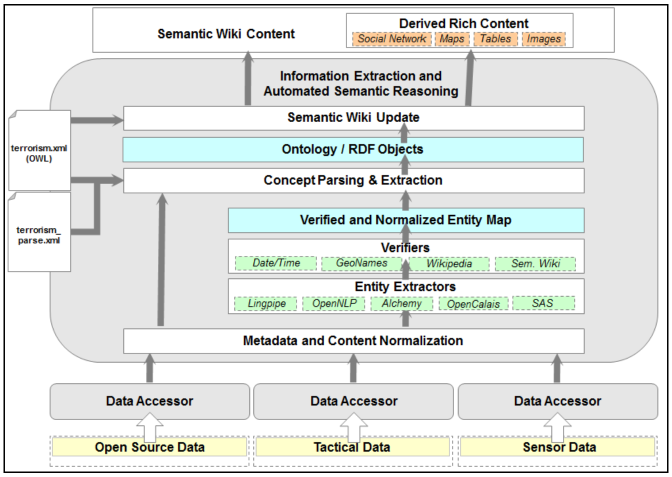


Figure 1: The Luminary Extraction Process. Luminary uses a multi-step process to extract and verify entities, and use those entities to identify more complex concepts.

1.1 Entity Extraction

Luminary incorporates interfaces to multiple entity extraction engines to improve accuracy. The system compares results from each extraction engine, and uses a best-guess approach to merging the results. The initial development of Luminary used Alchemy, Calais, Lingpipe and OpenNLP for entity extraction.

Different extraction engines use different algorithms to identify entities within a document, their performance and accuracy will vary. An extraction engine that performs well when parsing one set of document may perform poorly on another document corpus. Furthermore, those extraction engines that assign a type to the extracted entities frequently use different categorization schemes to

ed.swing@vsticorp.com

tag entities. By incorporating multiple entity extraction engines, Luminary mitigates the individual variances of individual engines. Luminary assembles the results, examines the different categories, and decides on the most appropriate entity type for each entity.

1.2 Entity Verification

Following the entity extraction process, the entities pass through a set of Entity Verifiers. These verifiers attempt to determine whether a particular entity is valid, enhancing the accuracy of the extraction process. To perform this process, the verifiers use external information sources. For instance, to determine if an entity is a person, the system checks the name on Wikipedia, and validates the entity if the page mentions a birth date.

At the same time, the verifiers also identify alternate spellings or references to that person, and normalize the entity. This ensures that a particular concept has a consistent representation, not only within a single document, but across all documents in the system. For instance, it could correctly determine that the expressions “Barack Obama” and “President Obama” refer to the same person when mentioned in a news article.

During this step, the Luminary system also retrieves any information on the entities from the Luminary wiki, and builds ontological objects for each entity.

1.3 Concept Extraction

The final step in the extraction process involved semantic concept extraction. During this step, Luminary attempts to extract composite concepts, such as a terrorist attack. This step identifies the concepts, and discovers properties for each concept. The concept extraction conforms to a specific OWL ontology. For the VAST challenge, we applied and augmented a terrorism ontology.

The rules used in the Luminary extraction process define not only the concept, but also its properties and how the concept relates to other objects. For instance, while many systems can identify a sentence as referring to a terrorist attack, they do not build a complex object that contains additional information such as the location, date, or responsible party. The Luminary process identifies this information using the conceptual rules, thus allowing for more complex analysis.

Luminary’s parsing rules incorporate typed variable substitution, allowing them to match particular entities. They also include stemming and the ability to recognize parts of speech. Using this, a rule expression might indicate that only a noun phrase might match a particular variable, allowing Luminary to discover unknown entities and concepts while parsing. The expressions within the rules can also include synonyms, using the Princeton WordNet.

For instance, when parsing the sentence “*Al-Qaeda militants claimed responsibility for an attack that killed 56 people in Kabul on Tuesday*” with the terrorism ontology, Luminary’s entity extraction and verification steps would identify Al-Qaeda as a TerroristGroup, which is a subclass of Organization. They would identify Kabul as a city, a subclass of Location, and Tuesday as a date-time reference. By normalizing the references (and assuming the text was from a news article written on August 20th, 2011), “Tuesday” would normalize to August 16th, 2011.

The lexical parsing rules would determine that this sentence describes a TerrorAttack. The rules would set the hasLocation property for the TerrorAttack object to Kabul, and the eventDate property to August 16th. It would identify Al-Qaeda as the actor, and set the numberKilled property to 56. The Luminary process assembles this information into an ontology object, and then adds that information into the semantic wiki. Luminary creates a new page for the new event, and updates other pages appropriately.

Current research for Luminary include early designs for incorporating ontological reasoning and inferencing, merging the Luminary rules structure with the SAS Enterprise Content Categorization product, and automatic rules generation.

2 SEMANTIC WIKIS

Wikis have become a common, almost ubiquitous, paradigm for storing topical information on the internet. Wikipedia, launched in 2001, demonstrates how wikis can provide a central reference and representation for knowledge. Wikipedia has millions of articles on almost every topic in over 250 languages. In addition to Wikipedia itself, topical wikis focusing on specific domains of interest have flourished.

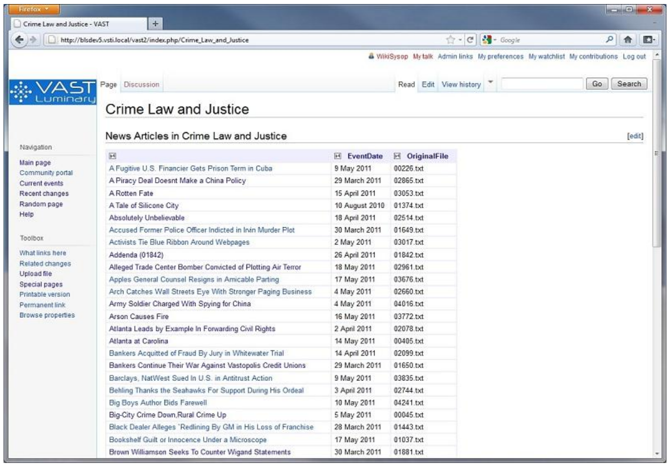
Semantic wikis have recently developed from advances in semantic web applications. Enhancing wikis with semantic constructs, wiki editors have developed many sites that contain semantically-rich information, presenting it in a compelling manner. Semantic wikis begin to blend the ubiquitous knowledge representation of a wiki with some of the knowledge formalism explored in ontological research. The editors of some semantic wikis, such as IkeWiki [7], have explored how to engineer the knowledge within semantic wikis for knowledge representation.

Luminary uses MediaWiki with the Semantic Bundle [8]. It also uses other semantic extensions as appropriate for the wiki, such as the Semantic Forms and Semantic Results Format extensions.

2.1 Semantic Properties

In a semantic wiki, each link from one wiki page to another may have a property attached to it. This simple concept enables a semantic wiki to incorporate some very powerful features. By including a property, each link on a page forms a natural RDF triple. The properties can refer to other pages, numeric values, dates, or other data types.

Semantic wikis allow pages to embed queries or searches on a particular property. The wiki can format the results of the query as a table or even use a visual representation. For instance, each of the news articles in the 2011 VAST challenge had a **hasTopic** property added to them. The wiki can then create a page listing all articles in a particular topic dynamically, as seen in Figure 2.



Crime Law and Justice

News Articles in Crime Law and Justice [edit]

	EventDate	OriginalFile
A Fugitive U.S. Financier Gets Prison Term in Cuba	9 May 2011	00226.txt
A Piracy Deal Doesnt Make a China Policy	29 March 2011	02865.txt
A Rotten Fate	15 April 2011	03053.txt
A Tale of Silicone City	10 August 2010	01374.txt
Absolutely Unbelievable	18 April 2011	02514.txt
Accused Former Police Officer Indicted in Irvin Murder Plot	30 March 2011	01649.txt
Activists Tie Blue Ribbon Around Webpages	2 May 2011	03017.txt
Addenda (01842)	26 April 2011	01842.txt
Alleged Trade Center Bomber Convicted of Plotting Air Terror	18 May 2011	02961.txt
Apples General Counsel Resigns in Amicable Parting	17 May 2011	03676.txt
Arch Catches Wall Streets Eye With Stronger Paging Business	4 May 2011	02660.txt
Army Soldier Charged With Spying for China	4 May 2011	04016.txt
Arson Causes Fire	16 May 2011	03772.txt
Atlanta Leads by Example In Forwarding Civil Rights	2 April 2011	02078.txt
Atlanta at Carolina	14 May 2011	00405.txt
Bankers Acquitted of Fraud By Jury in Whitewater Trial	14 April 2011	02099.txt
Bankers Continue Their War Against Vastopolis Credit Unions	29 March 2011	01650.txt
Barclays, NatWest Sued In U.S. in Antitrust Action	9 May 2011	03835.txt
Behling Thanks the Seahawks For Support During His Ordeal	3 April 2011	02744.txt
Big Boys Author Bids Farewell	10 May 2011	04241.txt
Big-City Crime Down,Rural Crime Up	5 May 2011	00045.txt
Black Dealer Alleges `Redlining By GM in His Loss of Franchise	28 March 2011	01443.txt
Bookshelf Guilt or Innocence Under a Microscope	17 May 2011	01037.txt
Brown Williamson Seeks To Counter Wigand Statements	30 March 2011	01881.txt

Figure 2: List of articles in the Crime Law and Justice topic, dynamically created by querying the hasTopic property in the news article pages.

This simple capability allows a semantic wiki to represent not only links between pages, but also the relationships between the objects represented by those pages.

All wikis also include the ability to assign pages to one or more categories. Most also allow a page to incorporate templates, which format text or embed common elements. For instance, many pages on Wikipedia use the Infobox template to encapsulate factual information in a small information box, typically displayed in the upper right corner of the page.

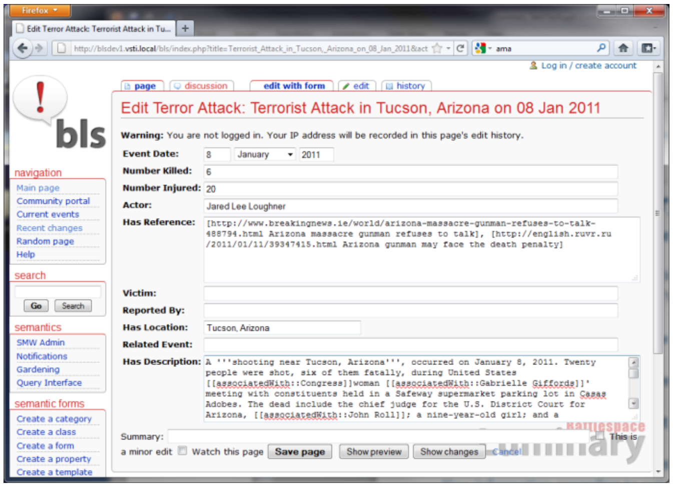
2.2 User interface Design

The contents of a traditional wiki page consist mostly of unstructured text with some template invocations. This approach is valid for unstructured concepts and a standalone wiki. However, when incorporating the wiki into an automated process, the pages require more structure.

Luminary approaches the semantic wiki framework differently from most semantic wikis. Pages within a Luminary Semantic Wiki consist entirely of a template invocation, and do not contain any free text common to most wikis. Instead, specific fields within the template invocation contain the unstructured descriptive text. This design allows a wiki editor to create templates that control the display of the entire content of the wiki page.

The design also enables an external process to query and update the information in the wiki while retaining its semantic structure. Using the Mediawiki API, Luminary can query the full contents of pages within the wiki. Parsing the contents of a page within a normal wiki would require natural language processing, and scanning the content for relevant semantic properties. Instead, with the Luminary wiki structure, the entire content of the web page falls into different semantic properties. Using this information, Luminary creates ontological objects from web pages and can update them consistently.

To the average user browsing wiki pages, this change is not apparent. For users who wish to edit wiki pages, the semantic wiki incorporates the Semantic Forms extension. A semantic form allows a user to edit the contents of a template. Therefore, within a Luminary wiki, this provides full page editing while retaining the structure of the page.



Edit Terror Attack: Terrorist Attack in Tucson, Arizona on 08 Jan 2011

Warning: You are not logged in. Your IP address will be recorded in this page's edit history.

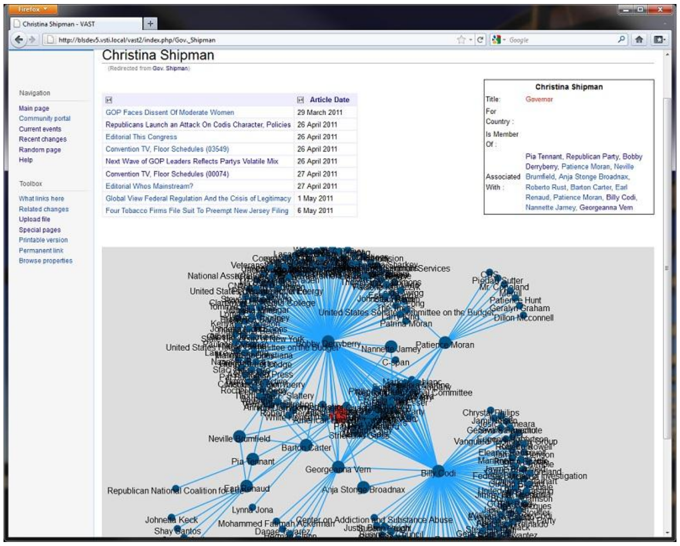
Event Date:	8 January 2011
Number Killed:	6
Number Injured:	20
Actor:	Jared Lee Loughner
Has Reference:	[http://www.breakingnews.ie/world/arizona-massacre-gunman-refuses-to-talk-488794.html Arizona massacre gunman refuses to talk], [http://english.ruvr.ru/2011/01/11/39347415.html Arizona gunman may face the death penalty]
Victim:	
Reported By:	
Has Location:	Tucson, Arizona
Related Event:	
Has Description:	A '''shooting near Tucson, Arizona''', occurred on January 8, 2011. Twenty people were shot, six of them fatally, during United States [[associatedWith::Congress]]woman [[associatedWith::Gabrielle Giffords]]' meeting with constituents held in a Safeway supermarket parking lot in Casas Adobes. The dead include the chief judge for the U.S. District Court for Arizona, [[associatedWith::John Roll]]; a nine-year-old girl; and a

Summary: This is a minor edit ☐ Watch this page Save page Show preview Show changes Cancel

Figure 3: Form entry for a terror attack. Fields correspond to properties in the wiki and the ontology, allowing an analyst to update semantic information without needing to understand wikitext.

2.3 Rich Content Visualizations

In a semantic wiki, the results of a semantic query can generate visualizations as well as tables or lists of pages. For instance, the semantic query within a wiki page can easily generate simple charts. A semantic query can also generate a timeline by including a date-time property. It can also display the results in a graph, providing social network displays, as seen in Figure 4.



Christina Shipman

(Redirected from Gov. Shipman)

	Article Date
GOP Faces Dissent Of Moderate Women	29 March 2011
Republicans Launch an Attack On Codis Character, Policies	26 April 2011
Editorial This Congress	26 April 2011
Convention TV, Floor Schedules (03549)	26 April 2011
Next Wave of GOP Leaders Reflects Partys Volatile Mix	26 April 2011
Convention TV, Floor Schedules (00074)	27 April 2011
Editorial Whos Mainstream?	27 April 2011
Global View Federal Regulation And the Crisis of Legitimacy	1 May 2011
Four Tobacco Firms File Suit To Preempt New Jersey Filing	6 May 2011

Christina Shipman

Title:	Governor
For Country :	
Is Member Of :	
Associated With :	Pia Tennant, Republican Party, Bobby Derryberry, Patience Moran, Neville Brumfield, Anja Stonge Broadnax, Roberto Rust, Barton Carter, Earl Renaud, Patience Moran, Billy Codi, Nannette Jamey, Georgeanna Vern

Figure 4: Page for a fictitious politician from the 2011 VAST Challenge, displaying a force-directed social network. The wiki creates the network by executing a query on the associatedWith property, arranging the results as a graph.

3 ONTOLOGY COORDINATION

To ensure consistency in both the content extraction and the semantic wiki contents, the Luminary system uses an OWL ontology to define and constrain the data of interest. This ontology represents the domain of interest for the analysis, and therefore the contents of the resulting wiki.

3.1 Conceptual Parsing with an Ontology

Many systems use parsing rules to identify concepts. However, Luminary's design incorporates the OWL ontology in the parsing process. Because of this, the semantic parsing rules not only identify concepts, but also identify related concepts and properties.

The rules for the concept extraction correspond to classes within the ontology. The ontology classes have sets of associated rule patterns that identify common ways to express a concept corresponding to that class. Just as classes within an ontology can have subclasses, rules for a subclass of objects can inherit rules from the parent class. For instance, the rules for SuicideBombing in the terrorism ontology incorporate the rules for TerrorAttack, since SuicideBombing extends TerrorAttack in the ontology.

Each of the ontology classes also specifies its valid properties. The properties can specify their type or ontology class. This allows the rules to determine when a parsing rule has discovered an appropriate value, so the rules can identify and eliminate improper values for properties.

3.2 Ontology and Wiki Design

The Luminary wiki design shifts wiki categories away from being simple tags to using categories and their corresponding templates to define the structure of information in the wiki. Luminary uses the OWL ontology to configure the wiki, and coordinates the templates, categories and properties to conform to the ontology. The category hierarchy within the wiki mimics the class hierarchy within the OWL ontology. Pages within the wiki are semantic objects conforming to the ontology. Properties in the ontology match the properties in the wiki.

Each category in the wiki may also have its own template, or may invoke the template of a parent class. Field names in the template match the properties in the corresponding OWL class, removing any challenge of automatically parsing information in the wiki pages. This template also handles embedding any queries and generating visualizations for all pages within the category. The template handles visual style and markup, so all pages within a category have a consistent layout and appearance.

4 Using Luminary for Text Analysis

The Luminary design makes it ideal for building and maintaining a collection of information about specific people, events, or organizations. It can help an analyst transform incoming news reports and other information into a cohesive repository of information. Luminary uses the wiki pages to represent the current knowledge about an entity or event – "what we know". The domain ontology represents the domain of information – "what we want to know".

The automated analysis process within Luminary allows analysts to tweak the lexical parsing rules for extracting concepts. An analyst can modify the conceptual extraction rules through an innovative rule builder that allows a user to highlight text and receive optional rule suggestions for a particular phrase or sentence.

Within the wiki, an analyst has several techniques they might use for analysis. All wikis provide simple word search capability, but a semantic wiki also allows an analyst to form semantic queries. For instance, an analyst might request a list of all wiki pages about terror attacks in Kabul in August 2011, displaying the results in a timeline display. Analysts can save these queries as wiki pages, thus allowing the analyst to share the results of his search with others. Furthermore, such queries are dynamic, so any time an analyst looks at the query page, the wiki will provide current results.

Other features of semantic wikis also support text analytic needs. Wikis include redirect pages that redirect a user to the correct page for a topic. For instance, the Wikipedia page for "Obama" redirects users to "Barack Obama". This mechanism provides a natural mechanism for entity co-reference resolution. The Luminary extraction process uses the semantic wiki to identify co-references automatically, and also updates the wiki when it identifies new co-reference terms.

No text extraction process performs flawlessly. Therefore, a text analysis system must allow analysts to modify and update any results generated by an automated process. The forms within the semantic wiki enable a user to edit the results of the automated analysis easily, fixing errors or adding new information.

Analysts must also collaborate with one another. They must also maintain a record of their analysis, or demonstrate where a particular nugget of information originated. All wikis incorporate a discussion page that allows an analyst to annotate a page with comments. Analysts can use the discussion page to collaborate with other analysts about the topic of the page, sharing their insights or debating the analysis. Finally, a wiki also allows an analyst to receive notifications when someone modifies a particular page or group of pages.

5 Conclusion

The Luminary prototype system offers a broad range of capabilities for analysts. Its growing collection of text analysis techniques can provide automated analysis on large collections of documents. Utilities for refining and focusing the process enable analysts to improve the automated parsing. By representing news articles and entities as ontological objects, Luminary can utilize sophisticated reasoning approaches to enhance traditional text analytics.

By using a semantic wiki as both the knowledge repository and the representation for the analytical information, the Luminary system provides analysts with a familiar paradigm for knowledge representation. Analysts can explore the results of their textual analysis in novel ways by employing the semantic capabilities within the wiki.

Using the Luminary approach, a semantic wiki can not only provide a compelling interface, but also integrate with other systems, and provide a rich experience for users. By combining and constraining both the extraction process and the contents of the wiki through an ontology, Luminary provides a compelling and consistent system for investigative analysis, collaboration, and reporting.

Acknowledgments

The Luminary system was initially developed under SBIR OSD09-SP6, "Semantic Wiki for Page Alerting" through the Office of Naval Research. Components of Luminary were also used in SBIR AF093-042, "Persistent Queries for Evolving Situational Awareness of Organization Entities."

References

[1] E. Agichtein, E, Eskin and L. Gravang, "Combining Strategies for Extracting Relations from Text Collections", Proceedings of the ACM SIGMOD Workshop on Data Mining and Knowledge Discovery (DMKD'00) (2000)

[2] E. Gabrilovich and S. Markovitch, "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge", proceedings of the 21st national conference on Artificial intelligence - Volume 2, AAAI press, 2006.

[3] ZGrimes, S., "Unstructured Data and the 80 Percent Rule", http://www.clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551

[4] N. Guarino, "Formal Ontology, Conceptual Analysis and Knowledge Representation", International Journal of Human-Computer Studies, Academic Press, Volume 43 Issue 5-6, Nov./Dec. 1995.

[5] M. Hepp, "OntoWiki: community-driven ontology engineering and ontology usage based on Wikis", Proceedings of the 2006 international symposium on Wikis, ACM, 2006, doi:10.1145/1149453.1149487

[6] M. Kokar, C. Matheus, and K. Baclawski, "Ontology-based situation awareness", Information Fusion, Volume 10, Issue 1, Special Issue on High-level Information Fusion and Situation Awareness, January 2009, Pages 83-98, ISSN 1566-2535, DOI: 10.1016/j.inffus.2007.01.004.

[7] S. Schaffert, "IkeWiki: A Semantic Wiki for Collaborative Knowledge Management", Enabling Technologies: Infrastructure for Collaborative Enterprises, June 2006. 15th IEEE International Workshops on Collaboration Technologies and Infrastructure, doi:10.1109/WETICE.2006.46.

[8] Semantic MediaWiki (website): http://semantic-mediawiki.org/wiki/Semantic_MediaWiki