# Text understanding agents and the Semantic Web

Akshay Java, Tim Finin and Sergei Nirenburg
University of Maryland, Baltimore County, Baltimore MD 21250
{aks1,finin,sergei}@umbc.edu

***Abstract*— We discuss the challenges involved in adapting the OntoSem natural language processing system to the Web. One set of tasks involves processing Web documents, translating their computed meaning representations from the OntoSem's native KR language into the Semantic Web language OWL, and publishing the results as Web pages and RSS feeds. Another set of tasks works in reverse – querying the Web for facts needed by OntoSem, translating them from OWL into OntoSem's native KR language and importing the results. A central problem underlying both sets of tasks is that of translating knowledge between OntoSem's KR language and ontologies and those of the Semantic Web. OntoSem2OWL has been developed as a translation system to support these translations. We describe SemNews, an implemented prototype application that demonstrates the process. It monitors RSS feeds of news stories, applies OntoSem to understand the text, and exports the computed facts back to the Web in OWL.**

## I. Introduction

The web has quickly grown from a modest hypertext system of interest to computer researchers to a ubiquitous information system including virtually all of human knowledge. Today's Web provides ready access to not only text, images, and audio files, but also to structured data, semi-structured information, services and people. It offers an open, decentralized (and uncontrollable!) environment in which anyone can publish information and services coupled with powerful search engines to find and rank relevant information and services. All of this is ubiquitously available from wired, wireless and mobile devices. Oh, and did we mention that it's free?

The result is an environment enormously useful to people for research, learning, commerce, socializing, communication and entertainment. For many people today, the first and sometimes only, resource used to answer a question, find a fact, or learn about topic is a quick search on the Web, mediated by a search engine such as Google, to find the most relevant documents needed. While the Web has made us all "smarter" by putting such information "at our fingertips", we have just begun to explore how this vast amount of machine accessible knowledge can be exploited and used by machines – to better serve human needs, to discover new knowledge and to acquire facts and knowledge essential to understanding text in a dynamic world.

Intelligent software agents need knowledge, information and data to perform their tasks. While some web information is directly encoded in forms that are relatively easy for agents to understand, such as XML or RDF, the vast majority is presented as natural language text. We anticipate a future in which sophisticated text understanding systems will process text found on the web and publish the results of their analyses on the web in a form accessible to other agents. One such form is as documents and annotations encoded in Semantic Web languages such as RDF and OWL. This will make the vast amount of information found in text documents on the Web more readily and efficiently available to a large community of software agents.

At the same time, language understanding agents can directly use information found on the web encoded in OWL to help guide their language understanding tasks. NLP systems require not only ontological knowledge (e.g., "A city is a geopolitical region") and lexical knowledge (e.g., "$city$ denotes a city") but also a considerable body of facts. Such information, often called encyclopedic knowledge, includes facts like "Colin Powel is the name of the current U.S. Secretary of State", "Annapolis is the name of the capital city of the State of Maryland" and "Colin Powel's boss is George Bush". While the Semantic Web contains ontological and lexical knowledge, it is an especially rich and useful source of facts.

We envision NLP agents turning to the Web to find information as they process text just as a human reader might. For example, when an NLP agent encounters the string *Michael Chertoff* in a news story, it can recognize that it is probably a name from various lexical and syntactic clues. Understanding who *Michael Chertoff* is could be important to fully understanding the rest of the text. So, our agent might query a Semantic Web search engine such as Swoogle [13] to find relevant facts about individuals with that name. The results, after being filtered to remove untrusted sources, can be translated from the original RDF representation and ontologies into a form understandable by the NLP system. Thus, the NLP system comes to learn that *Chertoff* is the head of the U.S. Department of Homeland Security, that he was born in Elizabeth NJ in 1953, that he is a registered Republican, etc.

This paper describes our initial work in exploring these ideas by adapting the OntoSem natural language processing system to the Web. One set of tasks involves processing Web documents, translating their computed meaning representations from the OntoSem's native KR language into the Semantic Web language OWL, and publishing the results as Web pages and RSS feeds. Another set of tasks works in reverse – querying the Web for facts needed by OntoSem, translating them from OWL into OntoSem's native KR language and importing the results. A central and challenging problem underlying both sets of tasks is that of translating knowledge between OntoSem's KR language and ontologies and those of the Semantic Web. In order to explore these tasks concretely, we have developed SemNews, an implemented prototype application that monitors RSS feeds of news stories, applies OntoSem to understand the text, and exports the computed facts back to the Web in OWL.

The remainder of this paper is organized as follows. We start with a brief review of some related work on mapping knowledge between a text understanding system and the Semantic Web representation. Section III provides an overview of the architecture of our implemented system and describes the approach used and major issues discovered in using it to map knowledge between OntoSem and OWL. Section IV outlines some of the larger issues and challenges we expect to encounter. While this work is still in a preliminary stage, we offer some thoughts on how some components can be evaluated in section V. Section VI describes the SemNews application testbed and VII describes some general application scenarios we have explored to motivate and guide our research. Finally, we offer some concluding remarks in section VIII.

## II. Related Work

Related work can be divided into three categories: systems that use language understanding techniques to extract information from Web resources; systems that translate knowledge and facts between different knowledge representation languages; and language understanding systems that dynamically import Semantic Web knowledge to use in language processing tasks.

Considerable work has been done on systems that extract information from text found on the web and represent it in a structured or semistructured formalism. Most apply simple information extraction techniques rather than the kind of deeper semantic analysis of which OntoSem is capable. Information extraction tools work best when the types of objects that need to be identified are clearly defined, for example the objective in MUC [17] was to find the various named entities in text. Using OntoSem, we aim to not only to provide such information, but also convert the text meaning representation of natural language sentences into Semantic Web representations.

The TAP [32] project is an example of a system that uses simple information extraction technologies to recognize named entities and simple relationships in Web text. The results are represented in RDF and supported by a shallow but broad knowledge base containing basic lexical and taxonomic information about a wide range of popular objects. TAP's focus on web-scale applications has meant that the language processing it can afford to do is quite limited.

Kruger et al. [24] developed an application that learned to extract information from talk and seminar announcements on the web from training data using an algorithm based on Stalker [27]. This system is typical of approaches that rely largely on machine learning techniques and use little or no language understanding technology. The extracted information was encoded as markup in DAML+OIL, a precursor to OWL, and used as part of the ITTALKS system [10].

An example of another approach is a system developed by the Haystack Project [19]. This semi-automated system enabled users to train a browser to extract Semantic Web content from HTML documents. Users highlight examples of semantic content and describing their desired meaning. Generalized wrappers are then constructed to extract information and encode the results in RDF. The goal is to let individual users generate Semantic Web content from text on web pages of interest to them.

There is a long history of work involving translation content from one knowledge representation language to another. Most relevant here is work that maps information between a frame-based KR system (like OntoSem's) and description logic representation system (like OWL).

A project closely related to our work was an effort to map the Mikrokosmos knowledge base to OWL [8], [9]. Mikrokosmos is a precursor to OntoSem and was developed with the original idea of using it as an interlingua in machine translation related work. This project developed some basic mapping functions that can create the class hierarchy and specify the properties and their respective domains and ranges. In our system we describe how facets, numeric attribute ranges can be handled and more importantly we describe a technique for translating the sentences from their Text Meaning Representation to the corresponding OWL representation thereby providing semantically marked up Natural Language text for use by other agents.

Oliver et al. [11] describe an approach to translating the Foundational Model of Anatomy (FMA) ontology to OWL. FMA is a large ontology of the human anatomy that was defined using a frame-based knowledge representation language. Some of the challenges faced were the lack of equivalent OWL representations for some frame based constructs and scalability and computational issues with the current reasoners.

Schlangen et al. [33] describe a system that that combines a natural language processing system with Semantic Web technologies to support the content-based storage and retrieval of medical pathology reports. The NLP component was augmented with a background knowledge component consisting of a domain ontology represented in OWL. The result supported the extraction of domain specific information from natural language reports which was then mapped back into a Semantic Web representation.

The Cyc project has developed a very large knowledge base of common sense facts and reasoning capabilities. Recent efforts [34] include the development of tools for automatically annotating documents and exporting the knowledge in OWL. The authors also highlight the difficulties in exporting an expressive representation like CycL into OWL due to lack of equivalent constructs.

While some systems have been designed that make use of knowledge bases expressed in OWL, we know of none which dynamically query the Semantic Web to find facts as they are needed.

## III. Architecture

Ontological Semantics (OntoSem) is a theory of meaning in natural language text [29]. The OntoSem environment is a rich and extensive tool for extracting and representing meaning in a language independent way. The OntoSem system is used for a number of applications such as machine translation, question answering, information extraction and language generation. It is supported by a *constructed world model* [30] encoded as a rich ontology. The Ontology is represented as a directed

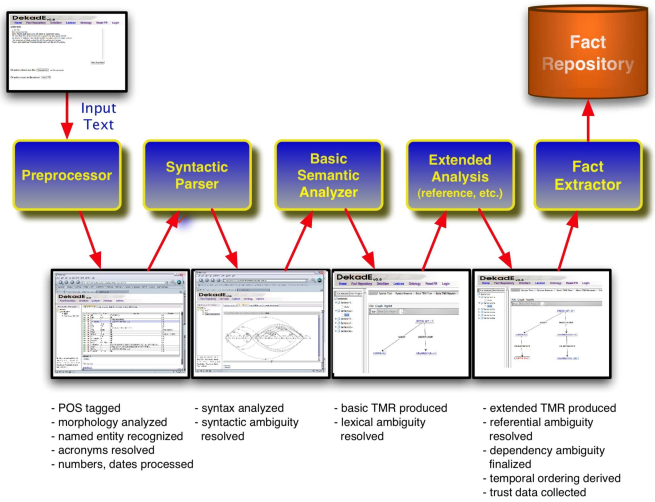

Input Text
Fact Repository
Preprocessor
Syntactic Parser
Basic Semantic Analyzer
Extended Analysis (reference, etc.)
Fact Extractor
- POS tagged
- morphology analyzed
- named entity recognized
- acronyms resolved
- numbers, dates processed
- syntax analyzed
- syntactic ambiguity resolved
- basic TMR produced
- lexical ambiguity resolved
- extended TMR produced
- referential ambiguity resolved
- dependency ambiguity finalized
- temporal ordering derived
- trust data collected

Fig. 1. OntoSem goes through several basic stages in converting a sentence into a text meaning representation (TMR).

acyclic graph using IS-A relations. It contains about 8000 concepts that have on an average 16 properties per concept. At the topmost level the concepts are: OBJECT, EVENT and PROPERTY.

The OntoSem ontology is expressed in a frame-based representation and each of the frames corresponds to a concept. The concepts are defined using a collection of slots that could be linked using IS-A relations. A slot consists of a PROPERTY, FACET and a FILLER.

```
ONTOLOGY ::= CONCEPT+
CONCEPT  ::= ROOT | OBJECT-OR-EVENT
                  | PROPERTY
SLOT     ::= PROPERTY + FACET + FILLER
```

The ontology is also supported by an *Onomasticon* [30], which is a lexicon of proper names. The learned instances from the text are stored in a *fact repository* which essentially forms the knowledge base of OntoSem. A more detailed description of OntoSem and its features is available in [30] and [3].

The OntoSem environment takes as input unrestricted text and performs different syntactic and semantic processing steps to convert it into a set of Text Meaning Representations (TMR). The basic steps in processing the sentence to extract the meaning representation is show in figure 1. The preprocessor deals with identifying sentence and word boundaries, part of speech tagging, recognition of named entities and dates, etc. The syntactic analysis phase identifies the various clause level dependencies and grammatical constructs of the sentence. The TMR is a representation of the meaning of the text and is expressed using the various concepts defined in the ontology. The TMRs are produced as a result of semantic analysis which uses knowledge sources such as lexicon, onomasticon and fact repository to resolve ambiguities and time references. Once the TMRs are generated, OntoSem2OWL converts them to an equivalent OWL representation.

In converting the OntoSem Ontology to OWL, we are performing the following tasks:

- Translating the OntoSem ontology, which deals with mapping the semantics of OntoSem into a corresponding OWL version.
- Once the ontology is translated the sentences that use the ontology are syntactically converted.
- In addition OntoSem is also supported by a fact repository which is also mapped to OWL.

Classes, properties and facets are the important constructs that need to be translated when converting OntoSem's frame based ontology to its corresponding OWL version. New concepts are defined in OntoSem using *make-frame* and related to other concepts using the *is-a* relation. Each concept may also have a corresponding definition. OBJECT or EVENT are mapped to *owl:Class* while, PROPERTIES are mapped to *owl:ObjectProperty*. ONTOLOGY-SLOTS are special properties that are used to structure the ontology. These are also mapped to *owl:ObjectProperty*. Object definitions are created using *owl:Class* and the IS-A relation is mapped using *owl:subClassOf*. Definition property in OntoSem has the same function as *rdfs:label* and is mapped directly.

Whenever the level one parent of a concept is of the type PROPERTY it is translated to *owl:ObjectProperty*. Properties can also be linked to other properties using the IS-A relation. In case of properties, the IS-A relation maps to the *owl:subPropertyOf*. Most of the properties also contain the domain and the range slots. Domain defines the concepts to which the property can be applied and the ranges are the concepts that the property slot of an instance can have as fillers. OntoSem domains are converted to *rdfs:domain* and ranges are converted to *rdfs:range*. For some of the properties OntoSem also defines inverses using the INVERSE-OF relationship. It can be directly mapped to the *owl:inverseOf* relation. Numerical attribute ranges for properties are handled by using *xsd:restriction*.

One of the important features of OntoSem ontology is the use of facets. Facets are a way of restricting the the fillers that can be used for a particular slot. The most commonly used facets are SEM and VALUE that indicate the value that the filler can take. We can map them using *owl:Restriction* thus locally restricting the type of values a property can take. The RELAXABLE-TO facet facet indicates that the value for the filler can take a certain type. It is a way of specifying "typical violations". One way of handling RELAXABLE-TO is to add this information in an annotation and also add this to the classes present in the *owl:Restriction*. DEFAULT and DEFAULT-MEASURE indicate the typical values or the typical units of measurement for a particular property. There is no clear way to express defaults in OWL since it only supports monotonic reasoning and this is one of the issues that have been expressed for future extensions of OWL language [20]. The NOT facet specifies that certain values are not permitted in the filler of the slot in which this is defined. *NOT* facet can be handled using the *owl:disjointWith* feature.

Once the OntoSem ontology is converted into its corresponding OWL representation, we can now translate the text meaning representations into statements in OWL. In order to do this we can use the namespace defined as the OntoSem ontology and use the corresponding concepts to create the representation. In addition TMRs also contain certain triggers for ’meaning procedures’ such as TRIGGER-REFERENCE and SEEK-SPECIFICATION. These are actually procedural attachments and hence can not be directly mapped into the corresponding OWL versions.

A more detailed description of the translation rules and issues are described in [21].

## IV. Challenges

There are a number of challenges in trying to map a frame based system like OntoSem to OWL. This section discusses some of the important issues that pertain to mapping of any frame based system to web representation such as OWL.

One of the challenges in building such a system is to bridge the gap between the knowledge representation features that are used by natural language processing systems and Semantic Web technologies. Typically NLP systems such as OntoSem are supported by frame based representations to construct a model or ontology of the world. Such an ontology is then used to extract and represent meaning from natural language text. Since OntoSem is used for natural language processing applications, it has a way of expressing defaults and exceptions. However there is no clear way of mapping defaults to OWL since OWL does not support nonmonotonic reasoning and has an open world assumption.

Knowledge sharing is a critical factor to enable agents on the Semantic Web to use this information extracted from NL text or be able to provide information that can be used by NLP tools. This requires mapping across different ontologies and translating sentences from one representation to another. KQML [15] and KIF [16] were two such attempts that developed protocols to enable sharing of large scale *knowledge bases*. Our system maps the OntoSem ontology to OWL and thus makes the framework sharable with other agents on the web.

Ambiguity is also an issue when dealing with NL text. Human language can have ambiguity at both syntactic and semantic level. An example often discussed is *anaphora resolution*, which is the problem of identifying and resolving different references to the same named entity. OntoSem provides ways for handling such references and resolves these references, not just within a single document but across all the facts in its repository. This could have interesting applications in the Semantic Web domain, especially in resolving ambiguities in inherent in FOAF [2] descriptions and data.

While some of the basic mapping rules have been developed, more needs to be done to identifying and represent cardinalities, transitive, symmetric and inverse functional properties. These issues are being investigated.

There were also interesting challenges while mapping a large ontology such as OntoSem. Although we needed the capabilities of OWL Full to represent a more complete subset of OntoSem’s features, the result was too large for OWL Full reasoners to process. One suggestion is to build mappings at different levels of expressivity, for example we could have different versions of the OntoSem ontology for OWL Lite, DL and Full. Another approach would be to investigate the possibility of partitioning the ontology into different smaller ontologies.

OntoSem uses procedural attachments with concepts in the ontology and also in the TMRs. These are useful in performing tasks such as reference resolution, finding the relative time reference, etc. An important implication of the translation process is that currently it does not support any of these procedural attachments. It would be interesting to look into ways in which this information could be additionally incorporated either into the reasoner or the knowledge base of the agent itself.

## V. Preliminary Evaluation

There are several dimensions along which this research could be evaluated. Our translation model involves translating ontologies and instances (facts) in both directions: from OntoSem to an OWL version of the OntoSem Ontology and from the OWL version of OntoSem into OntoSem. For the translation to be truly useful, it should also involves the translation between the OWL version of OntoSem’s ontologies and facts and the ontologies in common use on the Semantic Web (e.g., FOAF [2], Dublin Core [26], OWL-S [7], OWL-time [18], etc.).

Since our current work has concentrated on the initial step of translating from OntoSem to OWL, we will enumerate some of the issues from that perspective. Translating in the opposite direction raises similar, though not identical, issues. The chief translation measures we have considered are as follows:

- **Syntactic correctness.** Does the translation produce syntactically correct RDF and OWL? The resulting documents can be checked with appropriate RDF and OWL validation systems.
- **Semantic validity.** Does the translation produce RDF and OWL that is semantically well formed? An RDF or OWL file can be syntactically valid yet contain errors that violate semantic constrains in the language. For example, an OWL class should not be disjoint with itself if it has any instances. Several OWL validation services make some semantic checks in addition to syntactic ones. A full semantic validity check is quite difficult and, to our knowledge, no system attempts one, even for decidable subsets of OWL.
- **Meaning preservation.** Is the meaning of the generated OWL representation identical to that of the OntoSem representation? This is a very difficult question to answer, or even to formulate, given the vast differences between the two knowledge representation systems. However, we can easily identify some constructs, such as defaults, that clearly can not be captured in OWL, leading to a loss of information and meaning when going from OntoSem to OWL.
- **Feature minimization.** OWL is a complex representation language, some of whose features make reasoning difficult. A number of levels of complexity can be identified (e.g., the OWL *species: Lite, DL and Full*). In general, we would like the translation service to not use a complex feature unless it is absolutely required. Doing so will reduce the complexity of reasoning with the generated ontology.

- **Translation complexity.** What are the speed and memory requirements of the translation. Since, in general, a translation might require reasoning, this could be an issue.

Since our project is still in an early stage, we report on some preliminary evaluation metrics covering the basic OntoSem to OWL translation.

OntoSem2OWL uses the Jena Semantic Web Framework [25] internally to build the OWL version of the Ontology. The ontologies generated were successfully validated using two automated RDF validators: the W3C's RDF Validation Service [4] and the WonderWeb OWL Ontology Validator [6].

There were a total of about 8000 concepts in the original OntoSem ontology of which 7747 were successfully translated. The total number of triples generated was just over 100,000. These triples included a number of blank nodes – RDF nodes representing objects without identifiers that are required due to RDF's low-level triple representation.

Because the generated ontologies required the use of the OWL's *union* and *inverseOf* features, the results fall in the *OWL full* class in terms of the the level of expressivity.

Using the Jena API it takes about 10-40 seconds to build the model, depending upon the reasoner employed. The computation of transitive closure and basic RDF Schema inferencing takes approximately ten seconds on a typical workstation. The OWL Micro reasoner takes about 40 seconds while OWL Full reasoner fails, possibly due to the large search space. The OntoSem ontology in its OWL representation can be successfully loaded into the SWOOP [22] OWL editor for browsing, editing and further validation.

Based on our preliminary results, we found that OntoSem2OWL is able to translate most of the OntoSem ontology into a form that is syntactically valid and, in so far as current validators can tell, free of semantic problems. Some of the OntoSem concepts, less tan four percent, could not be translated by the current system at all. We were able to identify many of the constructs that were translated with some meaning loss. Chief among them were the use of default values. However, these were used relatively sparingly in the OntoSem ontology. Were were not able, in general, to automatically detect other differences between the semantics of the original OntoSem ontology and it's OWL translation. This remains an open problem for further research.

## VI. An Application Testbed

One of the motivations for integrating language understanding agents into the Semantic Web is to enable applications to use the information published in free text along with other Semantic Web data. SemNews [5] is a semantic news service that monitors different RSS news feeds and provides structured representations of the meaning of news articles found in them. As new articles appear, SemNews extracts the summary from the RSS description and processes it with OntoSem. The resulting TMR is then converted into OWL.

The need for content syndication on the Web has led to the popularity of RSS and ATOM. Wider adoption of these technologies by content providers, blogging tools and news portals has also made available a number of aggregator tools and services such as Mozilla Thunderbird, Bloglines, my Yahoo Portal. RSS and Atom has also minimized the need to bookmark pages, enabling users to monitor news and other dynamic content by subscribing to the feeds. Another advantage of using RSS is the ability to provide text summaries that are manually or automatically generated.

Figure 2 shows the basic architecture of SemNews. The RSS feeds from different news sources are aggregated and parsed. These RSS feeds are also rich in useful meta-data such as information on the author, the date when the article was published, the news category and tag information. These form the explicit meta-data that is provided by the publisher. However there is a large portion of the RSS field that is essentially plain text and does not contain any semantics in them. It would be of great value if this text available in description and comment fields for example could be *semantacized*. By using Natural Language Processing (NLP) tools such as OntoSem we can convert natural language text into a structured representation thereby adding additional metadata in the RSS fields. Once processed, it is converted to its Text Meaning Representation (TMR). OntoSem also updates its fact repositories to store the information found in the sentences processed. These facts extracted help the system in its text analysis tasks.

An optional step of correction of the TMRs could be performed by means of the Dekade environment[1]. This is helpful in correcting cases where the analyzers are not able to correctly annotate parts of the sentence. Corrections can be performed at both the syntactic processor and the semantic analyzer phase. The Dekade environment could also be used to edit the OntoSem ontology and lexicons or static knowledge sources.

As discussed in the previous sections, the meaning in these structured representations, also known as Text Meaning Representations (TMR), can be preserved by mapping them to OWL/RDF. The OWL version of a document's TMRs is stored in a Redland-based triple store, allowing other applications and users to perform semantic queries over the documents. This enables them to search for information that would otherwise not be easy to find using simple keyword based search. The TMRs are also indexed by the Swoogle Semantic Web Search system [13].

The following are some examples of queries that go beyond simple keyword searches.

- **Conceptually searching for content.** Consider the query *"Find all stories that have something to do with a place and a terrorist activity"*. Here the goal is to find the content or the story, but essentially by means of using ontological concepts rather than string literals. So for example, since we are using the ontological concepts here, we could actually benefit from resolving different kinds of terror events such as bombing or hijacking to a terrorist-activity concept.
- **Context based querying.** Answering the query *"Find all the events in which 'George Bush' was a speaker"* involves finding the context and relation in which a particular concept occurs. Using named entity recognition alone, one can only find that there is a story about a named entity of the type person/human, however it is not

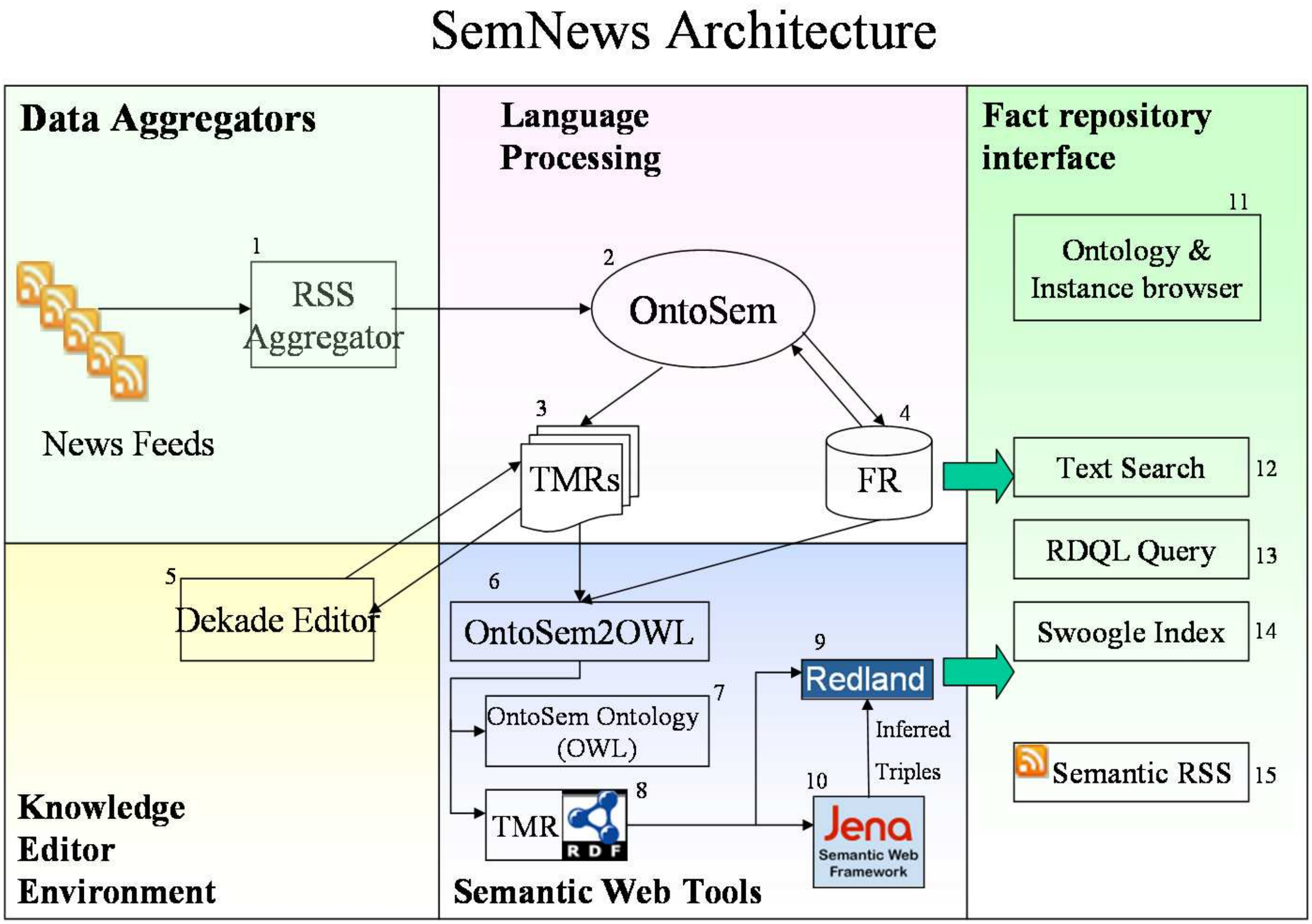


Fig. 2. The SemNews application, which serves as a testbed for our work, has a simple architecture. RSS (1) from multiple sources is aggregated and then processed by the OntoSem (2) text processing environment. This results in the generation of TMRs (3) and updates to the fact repository (4). The Dekade environment (5) can be used to edit the ontology and TMRs. OntoSem2OWL (6) converts the ontology and TMRs to their corresponding OWL versions (7,8). The TMRs are stored in the Redland triple store (9) and additional triples inferred by Jena (10). There are also multiple viewers for searching and browsing the fact repository and triple store.

directly perceivable as to what role the entity participated in. Since OntoSem uses deeper semantics, it not only identifies the various entities but also extracts the relations in which these entities or instances participate, thereby providing additional contextual information.

- **Reporting facts.** To answer a query like *”Find all politicians who traveled to ’Asia’”* requires reasoning about people’s roles and geography. Since we are using ontological concepts rather than plain text and we have certain relations like meronomy/part-of we could recognize that Colin Powel’s trip to China will yield an answer.
- **Knowledge sharing on the semantic web.** Knowledge sharing is critical for agents to reason on the semantic web. Knowledge can be shared by means of using a common ontology or by defining mappings between existing ontologies. One of the benefits of using a system like SemNews is that it provides a mechanism for agents to populate various ontologies with live and updated information. While FOAF has become a very popular mechanism to describe a person’s social network, not everyone on the web has a FOAF description. By linking the FOAF ontology to OntoSem’s ontology we could populate additional information and learn new instances of foaf:person even though these were not published explicitly in foaf files but as plain text descriptions in news articles.

The SemNews environment also provides a convenient way for the users to query and browse the fact repository and triple store. Figure 4 shows a view that lists the named entities found in the processed news summaries. Using an ontology viewer the user can navigate through the news stories conceptually while viewing the instances that were found. The fact repository explorer shown in Figures 5 and 6 provides a way to view the relations between different instances and see the news stories in which they were found. An advanced user may also query the triple store directly, using RDQL query language as shown in Figure 7. Additionally the system can also publish the RSS feed of the query results allowing users or agents to easily monitor new answers. This is a useful way of handling standing queries and finding news articles that satisfy a structured query.

Developing SemNews provided a perspective on some of the general problems of integrating a mature language processing system like OntoSem into a Semantic Web oriented application. While doing a complete and faithful translation of knowledge from OntoSem's native meaning representation language into OWL is not feasible, we found the problems to be manageable in practice for several reasons.

First, OntoSem's knowledge representation features that were most problematic for translation are not used with great frequency. For example, the default values, relaxable range constraints and procedural attachments were used relatively rarely in OntoSem's ontology. Thus shortcomings in the OWL version of OntoSem's ontology are limited and can be circumscribed. We are also optimistic that most Semantic Web content will be amenable to translation into OntoSem's representation. It's likely that the majority of Semantic Web content will be encoded with relatively simple ontologies that use only RDF and RDFS and do not use OWL. Many of the OWL ontologies may be partionable into portions which do not use difficult to translation features and those that do.

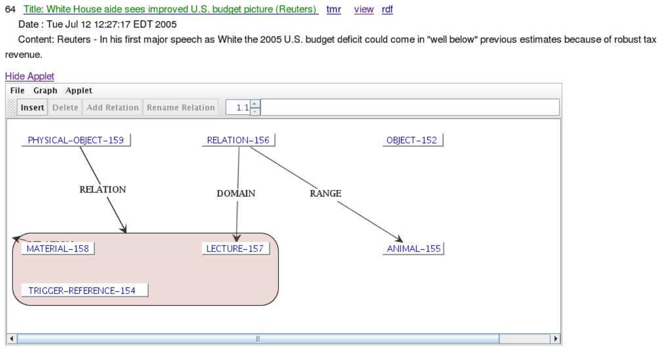
Second, the goal is not just to support translation between OntoSem and a complete and faithful OWL version of OntoSem. It is unlikely that most Semantic Web content producers or consumers will use OntoSem's ontology. Rather, we expect common consensus ontologies like FOAF, Dublin Core, and SOUPA to emerge and be widely used on the Semantic Web. The real goal is thus to mediate between OntoSem and a host of such consensus ontologies. We believe that these translations between OWL ontologies will of necessity be inexact and thus introduce some meaning loss or drift. So, the translation between OntoSem's native representation and the OWL form will not be the only lossy one in the chain.

Third, the SemNews application generates and exports facts, rather than concepts. The prospective applications coupling a language understanding agent and the Semantic Web that we have examined share this focus on importing and exporting instance level information. To some degree, this obviates many translation issues, since these mostly occur at the concept level. While we may not be able to exactly express OntoSem's complete concept of a book's author in the OWL version, we can translate the simple instance level assertion that a known individual is the author of a particular book and further translate this into the appropriate triple using the FOAF and Dublin Core RDF ontologies.

Finally, with a focus on importing and exporting instances and assertions of fact, we can require these to be generated using the native representation and reasoning system. Rather than exporting OntoSem's concept definitions and a handful of facts to OWL and then using an OWL reasoner to derive the additional facts which follow, we can require OntoSem to precompute all of the relevant facts. Similarly, when importing information from an OWL representation, the complete model can be generated and just the instances and assertions translated and imported.

## VII. Further Applications

Language understanding agents could not only empower Semantic Web applications but also create a space where humans and NLP tools would be able to make use of existing structured or semi structured information available. The following are a few of the example application scenarios.



64 Title: White House aide sees improved U.S. budget picture (Reuters) tmr view rdf

Date : Tue Jul 12 12:27:17 EDT 2005

Content: Reuters - In his first major speech as White the 2005 U.S. budget deficit could come in "well below" previous estimates because of robust tax revenue.

Fig. 3. A graphical view of the TMRs generated. TMRs are also exported in OWL.

### A. Semantic Annotation and Metadata Generation

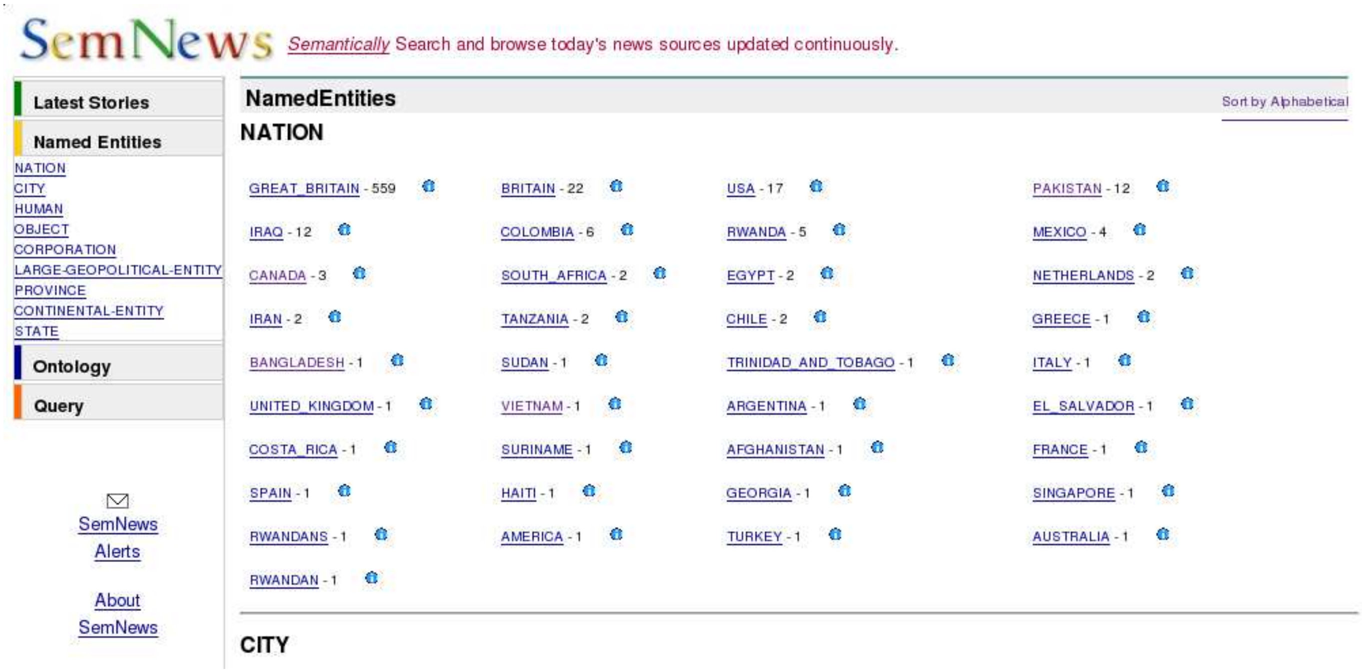
The growing popularity of folksonomies and social bookmarking tools such as del.icio.us have demonstrated that lightweight tagging systems are useful and practical. Metadata is also available in RSS and ATOM feeds, while some use the Dublin Core ontology. Some NLP and statistical tools such as SemTag[12] and the TAP[32] project aim to generate semantically annotated pages from already existing documents on the web. Using OntoSem in the SemNews framework we have been able to demonstrate the potential of large scale semantic annotation and automatic metadata generation. Figure 3 shows the graphical representation of the TMRs, which are also exported in OWL and stored in a triple store.

### B. Gathering Instances

Ontologies for the Semantic Web define the concepts and properties that the agents could use. By making use of these ontologies along with instance data agents can perform useful reasoning tasks. For example, an ontology could describe that a country is a subclass of a geopolitical entity and that a geopolitical entity is a subclass of a physical entity. Automatically generating instance data from natural language text and populating the ontologies could be an important application of such technologies. For example, in SemNews you can not only view the different named entities as shown in Figure 4 but also explore the facts found in different documents about that named entity. As shown in 5 and 6, we could start browsing from an instance of the entity type 'NATION' and explore the various facts that were found in the text about that entity. Since OntoSem also handles referential ambiguities, it would be able to identify that an instance described in one document is the same as the instance described in another document.

### C. Provenance and Trust

Provenance involves identifying source of information and tracking the history of where the information came from. Trust is a measure of the degree of confidence one has for a source



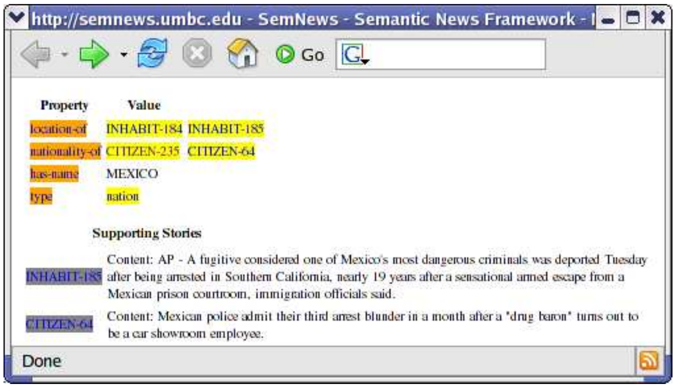
NamedEntities

NATION

| | | | |
|---|---|---|---|
| GREAT_BRITAIN - 559 | BRITAIN - 22 | USA - 17 | PAKISTAN - 12 |
| IRAQ - 12 | COLOMBIA - 6 | RWANDA - 5 | MEXICO - 4 |
| CANADA - 3 | SOUTH_AFRICA - 2 | EGYPT - 2 | NETHERLANDS - 2 |
| IRAN - 2 | TANZANIA - 2 | CHILE - 2 | GREECE - 1 |
| BANGLADESH - 1 | SUDAN - 1 | TRINIDAD_AND_TOBAGO - 1 | ITALY - 1 |
| UNITED_KINGDOM - 1 | VIETNAM - 1 | ARGENTINA - 1 | EL_SALVADOR - 1 |
| COSTA_RICA - 1 | SURINAME - 1 | AFGHANISTAN - 1 | FRANCE - 1 |
| SPAIN - 1 | HAITI - 1 | GEORGIA - 1 | SINGAPORE - 1 |
| RWANDANS - 1 | AMERICA - 1 | TURKEY - 1 | AUSTRALIA - 1 |
| RWANDAN - 1 | | | |

CITY

Fig. 4. Various types of named entities can be identified and explored in SemNews.


http://semnews.umbc.edu - SemNews - Semantic News Framework - I
Go
Done

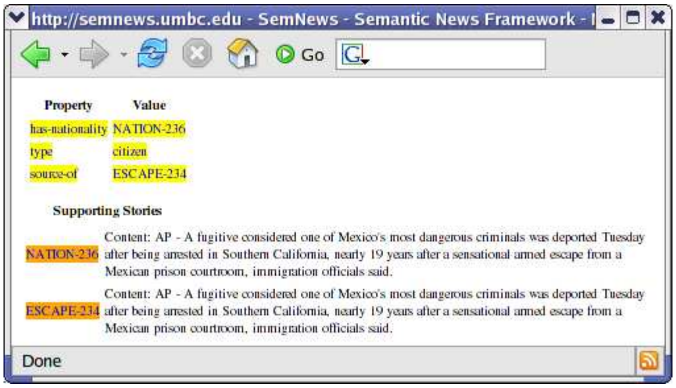
| Property | Value |
|---|---|
| location-of | INHABIT-184 INHABIT-185 |
| nationality-of | CITIZEN-235 CITIZEN-64 |
| has-name | MEXICO |
| type | nation |

Supporting Stories

INHABIT-185 Content: AP - A fugitive considered one of Mexico's most dangerous criminals was deported Tuesday after being arrested in Southern California, nearly 19 years after a sensational armed escape from a Mexican prison courtroom, immigration officials said.

CITIZEN-64 Content: Mexican police admit their third arrest blunder in a month after a "drug baron" turns out to be a car showroom employee.

Fig. 5. Fact repository explorer for the named entity ’Mexico’. Shows that the entity has a relation ’nationality-of’ with CITIZEN-235.


http://semnews.umbc.edu - SemNews - Semantic News Framework - I
Go
Done

| Property | Value |
|---|---|
| has-nationality | NATION-236 |
| type | citizen |
| source-of | ESCAPE-234 |

Supporting Stories

NATION-236 Content: AP - A fugitive considered one of Mexico's most dangerous criminals was deported Tuesday after being arrested in Southern California, nearly 19 years after a sensational armed escape from a Mexican prison courtroom, immigration officials said.

ESCAPE-234 Content: AP - A fugitive considered one of Mexico's most dangerous criminals was deported Tuesday after being arrested in Southern California, nearly 19 years after a sensational armed escape from a Mexican prison courtroom, immigration officials said.

Fig. 6. Fact repository explorer for the instance CITIZEN-235 shows that the citizen is an agent-of an ESCAPE-EVENT.

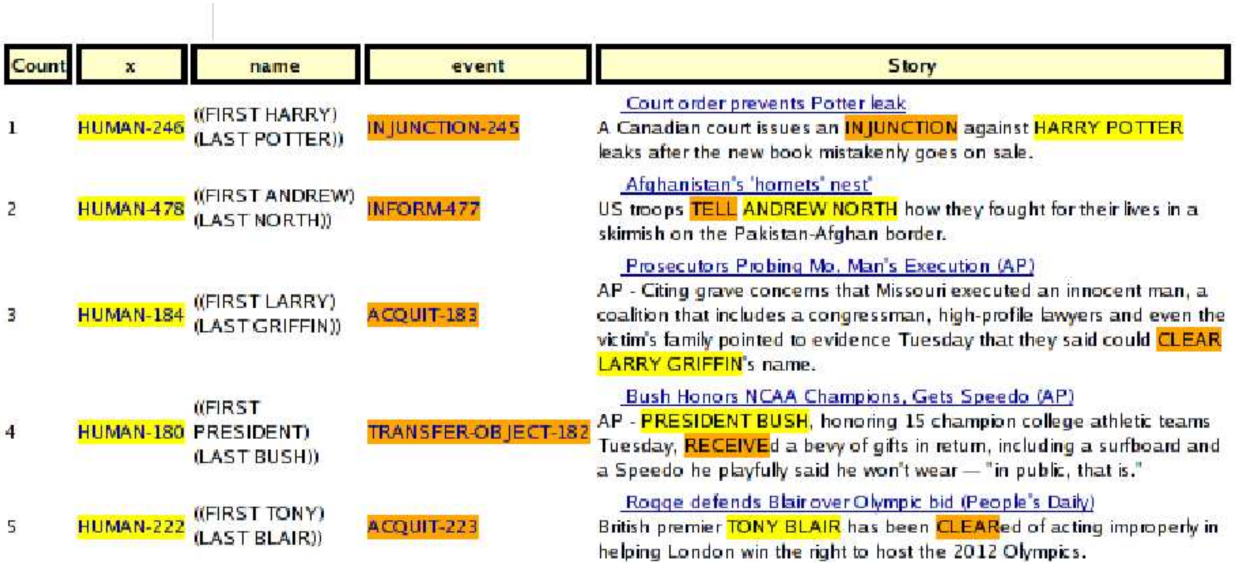
of information. While these are somewhat hard to quantify and are a function of a number of different parameters, there can be significant indicators of trust and provenance already present in the text and could be extracted by the agent. News report typically describe some of the provenance information as well as other metadata that can effect trust such as temporal information. This type of information would be important in applications where agents need to make decisions based on the validity of certain information.

### *D. Reasoning*

While currently reasoning on the Semantic Web is enabled by using the ontologies and Semantic Web documents, there could be potentially vast knowledge present in natural language. It would be useful to build knowledge bases that could not only reason based on explicit information available in them, but also use information extracted form natural language text to augment their reasoning. One of the implications of using the information extracted from natural language text in reasoning applications is that agents on the Semantic Web would need to reason in presence of inconsistent or incomplete annotations as well. Reasoning could be supported from not just semantic web data and natural language text but also based on provenance. Developing measures for provenance and trust would also help in deciding the degree of confidence that the reasoning engine may have in the using certain assertions for reasoning.

### *E. Ontology Enrichment*

Knowledge acquisition is one of the most expensive steps in developing large scale Semantic Web applications. Even within the framework of OntoSem, the OntoSem ontology has been developed and perfected over years of research in



| Count | x | name | event | Story |
|---|---|---|---|---|
| 1 | HUMAN-246 | ((FIRST HARRY)<br>(LAST POTTER)) | INJUNCTION-245 | Court order prevents Potter leak<br>A Canadian court issues an INJUNCTION against HARRY POTTER leaks after the new book mistakenly goes on sale. |
| 2 | HUMAN-478 | ((FIRST ANDREW)<br>(LAST NORTH)) | INFORM-477 | Afghanistan's 'hornets' nest'<br>US troops TELL ANDREW NORTH how they fought for their lives in a skirmish on the Pakistan-Afghan border. |
| 3 | HUMAN-184 | ((FIRST LARRY)<br>(LAST GRIFFIN)) | ACQUIT-183 | Prosecutors Probing Mo. Man's Execution (AP)<br>AP - Citing grave concerns that Missouri executed an innocent man, a coalition that includes a congressman, high-profile lawyers and even the victim's family pointed to evidence Tuesday that they said could CLEAR LARRY GRIFFIN's name. |
| 4 | HUMAN-180 | ((FIRST PRESIDENT)<br>(LAST BUSH)) | TRANSFER-OBJECT-182 | Bush Honors NCAA Champions, Gets Speedo (AP)<br>AP - PRESIDENT BUSH, honoring 15 champion college athletic teams Tuesday, RECEIVEd a bevy of gifts in return, including a surfboard and a Speedo he playfully said he won't wear — "in public, that is." |
| 5 | HUMAN-222 | ((FIRST TONY)<br>(LAST BLAIR)) | ACQUIT-223 | Rogge defends Blair over Olympic bid (People's Daily)<br>British premier TONY BLAIR has been CLEARed of acting improperly in helping London win the right to host the 2012 Olympics. |

Fig. 7. This SemNews interface shows the results for query “Find all humans and what are they the beneficiary-of”

linguistics, NLP and knowledge representation. In order to make the task of a knowledge engineer easier, we could possibly use the existing ontologies on the Semantic Web to suggest new concepts, relations or even properties. As an example consider the concept of fish, in OntoSem there are about 4 different varieties of fish that have been defined. We could now use a semantic search engine such as Swoogle [13] to find new types of fish and suggest some of the properties that could be used in order to describe fish in the ontology.

### F. Natural Language Interface to Semantic Web

While the Semantic Web is primarily for use by machines and the information available on it is in machine understandable format, the end goal is still to assist the human users in their tasks. Using technologies from question answering and language generation, it would be helpful to provide capabilities through which users can interact with their agent through natural language, thus reducing the cognitive load in formulating the task in a machine readable format.

## VIII. Conclusion

Natural language processing agents can provide a service by analyzing text documents on the Web and publishing Semantic Web annotations and documents that capture aspects of the text's meaning. Their output will enable many more agents to benefit from the knowledge and facts expressed in the text. Similarly, language processing agents need a wide variety of knowledge and facts to correctly understand the text they process. Much of the needed knowledge may be found on the Web already encoded in RDF and OWL and thus easy to import.

One of the key problems to be solved in order to integrate language understanding agents into the Semantic Web is translating knowledge and information from their native representation systems to Semantic Web languages. We have described initial work aimed at preparing the the OntoSem language understanding system to be integrated into applications on the Web. OntoSem is a large scale, sophisticated natural language understanding system that uses a custom frame-based knowledge representation system with an extensive ontology and lexicon. These have been developed over many years and are adapted to the special needs of text analysis and understanding.

We have described a translation system, OntoSem2OWL, that is being used to translate OntoSem's ontology into the Semantic Web language OWL. While the translator is not able to handle all of OntoSem's representational features, it is able to translate a large and useful subset. The translator has been used to develop SemNews as a prototype of a system that reads summaries of web news stories and publishes OntoSem's understanding of their meaning on the web encoded in OWL.

## References

[1] The dekade environment. `http://thoth.ilit.umbc.edu/servlet/dekade/index.jsp`.

[2] The friend of a friend(foaf) project. `http://www.foaf-project.org/`.

[3] Institute for language and information technologies. `http://ilit.umbc.edu/`.

[4] RDF validation service. `http://www.w3.org/RDF/Validator/`.

[5] Semnews application. `http://semnews.umbc.edu/`.

[6] Wonderweb owl ontology validator. `http://phoebus.cs.man.ac.uk:9999/OWL/Validator`.

[7] OWL web ontology language for services (OWL-S). A W3C submission, 2004. `http://www.w3.org/Submission/2004/07/`.

[8] P. Beltran-Ferruz, P. Gonzalez-Caler, and P.Gervas. Converting frames into OWL: Preparing Mikrokosmos for linguistic creativity. In *LREC Workshop on Language Resources for Linguistic Creativity*, 2004.

[9] P. Beltran-Ferruz, P. Gonzalez-Caler, and P.Gervas. Converting Mikrokosmos frames into description logics. In *RDF/RDFS and OWL in Language Technology: 4th ACL Workshop on NLP and XML*, July 2004.

[10] R. S. Cost, T. Finin, A. Joshi, Y. Peng, C. Nicholas, I. Soboroff, H. Chen, L. Kagal, F. Perich, Y. Zou, and S. Tolia. ITtalks: A Case Study in the Semantic Web and DAML+OIL. *IEEE Intelligent Systems Special Issue*, January 2002.

[11] O. Dameron, D. L. Rubin, and M. A. Musen. Challenges in converting frame-based ontology into OWL: the Foundational Model of Anatomy case-study. In *American Medical Informatics Association Conference AMIA05*, 2005.

[12] S. Dill, N. Eiron, D. Gibson, D. Gruhl, and R. Guha. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of WWW2003*, 2003.

[13] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi, , and J. Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, 2004.

[14] D. Dou, D. McDermott, and P. Qi. Ontology translation on the semantic web. *LNCS Journal of Data Semantics*, II(LNCS 3360):35–57, 2004.

[15] T. Finin, R. Fritzson, D. McKay, and R. McEntire. KQML as an Agent Communication Language. In N. Adam, B. Bhargava, and Y. Yesha, editors, *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94)*, pages 456–463, Gaithersburg, MD, USA, 1994. ACM Press.

[16] M. Genesereth. Knowledge interchange format. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning*, pages 599–600. Morgan Kaufmann, 1991.

[17] R. Grishman and B. Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 466–471, 1996.

[18] J. R. Hobbs and F. Pan. An ontology of time for the semantic web. *ACM Transactions on Asian Language Processing (TALIP)*, 3(1):66–85, March 2004. Special issue on Temporal Information Processing.

[19] A. Hogue and D. R. Karger. Thresher: Automating the unwrapping of semantic content from the world wide web. In *Proceedings of the Fourteenth International World Wide Web Conference*, May 2005.

[20] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From shiq and rdf to owl: the making of a web ontology language. *J. Web Sem.*, 1(1):7–26, 2003.

[21] A. Java, T. Finin, and S. Nirenburg. Integrating language understanding agents into the semantic web. In T. Payne and V. Tamma, editors, *Proceedings of the AAAI Fall Symposium on Agents and the Semantic Web*, November 2005.

[22] A. Kalyanpur, B. Parsia, and J. Hendler. A tool for working with web ontologies. In *In Proceedings of the International Journal on Semantic Web and Information Systems*, volume 1, January-March 2005.

[23] B. Katz, J. Lin, and D. Quan. Natural language annotations for the semantic web. In *International Conference on Ontologies, Databases and Applications of Semantics*, October 2002.

[24] W. Krueger, J. Nilsson, T. Oates, and T. Finin. *Automatically Generated DAML Markup for Semistructured Documents*. Lecture Notes in Artificial Intelligence. Springer, January 2004.

[25] B. McBride. Jena: Implementing the RDF model and syntax specification. In *Proceedings of the WWW2001 Semantic Web Workshop*, 2001.

[26] D. B. E. Miller and D. Brickley. Expressing simple dublin core in RDF / XML. Dublin Core Metadata Initiative Recommendation, 2002.

[27] I. A. Muslea, S. Minton, and C. Knoblock. Hierarchical wrapper induction for semistructured information services. *Journal of Autonomous Agents and Multi-Agent Systems*, 4(1/2):93–114, 2001.

[28] S. Nirenburg, S. Beale, and M. McShane. Evaluating the performance of the ontosem semantic analyzer. In *Proceedings of the ACL Workshop on Text Meaning Representation*, 2004.

[29] S. Nirenburg and V. Raskin. Ontological semantics, formal ontology, and ambiguity. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 151–161, New York, NY, USA, 2001. ACM Press.

[30] S. Nirenburg and V. Raskin. *Ontological semantics*. MIT Press, 2005.

[31] D. L. Rubin, O. Dameron, and M. A. Musen. Challenges in converting frame-based ontology into OWL: the foundational model of anatomy case-study. (submitted), 2005.

[32] R.V.Guha and R. McCool. TAP: A semantic web toolkit. *Semantic Web Journal*, October 2003.

[33] D. Schlangen, M. Stede, and E. P. Bontas. Feeding owl: Extracting and representing the content of pathology reports. In *RDF/RDFS and OWL in Language Technology: 4th ACL Workshop on NLP and XML*, July 2004.

[34] M. Witbrock, K. Panton, S. Reed, D. Schneider, B. Aldag, M. Reimers, and S. Bertolo. Automated OWL Annotation Assisted by a Large Knowledge Base. In *Workshop Notes of the 2004 Workshop on Knowledge Markup and Semantic Annotation at the 3rd International Semantic Web Conference ISWC2004*, November 2004.