

NEWS: Bringing Semantic Web Technologies into News Agencies

Norberto Fernández, José M. Blázquez, Jesús A. Fisteus, Luis Sánchez,
Michael Sintek, Ansgar Bernardi, Manuel Fuentes,
Angelo Marrara, and Zohar Ben-Asher

The NEWS Consortium*
info@news-project.com

Abstract. In the current Information Society, being informed is a basic necessity. As one of the main news business actors, news agencies are required to provide fresh, relevant, high-quality information to their customers. Dealing with this requirement is not an easy task, but, as partners of the NEWS (News Engine Web Services) project, we believe that the usage of Semantic Web technologies could help news agencies in achieving that objective. In this paper we will describe the aims and main achievements of the NEWS project, that was just completed.

1 Introduction

The news business is based on producing news items about current events and delivering them to customers. Customers want to receive information about events as soon as they occur. Customers do not want to be bothered with useless information, that is, they want to get information only about events of interest.

Dealing with both requirements, information freshness and relevance, is not easy to do. For instance, state of the art workflows at the news agencies try to address the relevance requirement by adding metadata to news items. These metadata are used to filter the news items according to the interests of each customer. But currently the metadata are added manually by the journalists and this is a time consuming task. Of course the negative effect of this metadata addition overhead can be reduced by restricting the metadata to be added to a few, easy to add, items: news item priority, keywords, editing location, category (economy, sports, ...), etc. This is the approach currently followed by the agencies. For instance, the categorisation is currently done based on small ($\simeq 10$ categories), in-house, category systems, easy to learn and use by the journalists. Finer grained annotations of the news items' contents (like basic entities occurrence) are not currently added. The effect of these restrictions on the set

* The NEWS Consortium is composed of news agencies *Agencia EFE S.A.* and *Agencia ANSA S.C.R.A.L.*, the *Deutsches Forschungszentrum für Künstliche Intelligenz GmbH* Research Institute, *Ontology Ltd.* and *Universidad Carlos III de Madrid* university.

of annotations to be added, is that high quality information filtering services are difficult to implement, which is opposite to the relevance requirement.

So, the current situation at news agencies as the Spanish EFE or the Italian ANSA is that customers access the news content either as a streaming according to their profiles (categories and other basic metadata), *push distribution mode*, or making free text queries over a news item repository, *pull distribution mode*. This of course results in receiving a lot of non-relevant content that the customer should manually filter. Additionally due to the fact that queries over news agencies' repositories are keyword-based, multilingual information retrieval is not feasible: the customer gets news items in the language of the terms (s)he has introduced in the free text query.

The NEWS [1] EU IST project aims at providing solutions which help news agencies to overcome limitations in their current workflows and increase their productiveness and revenues. In order to reach this aim, the NEWS project makes use of state-of-the-art Semantic Web technologies. In that sense, the work developed in this project covers mainly three topics:

Annotation: Implementing a semantic annotation component which automatically produces metadata annotations for news items. In the context of NEWS, the core of this semantic annotation component is the natural language processing (NLP) engine of Ontology Ltd.

Intelligent Information Retrieval: Developing intelligent components for the news domain, with multilingual and multimedia capabilities, which use semantic annotations and ontologies to allow the development of intelligent services for the news domain. In the context of NEWS we have developed a Heuristic and Deductive Database (HDDDB) component and the NEWS Ontology [2], which covers the main concepts required in the news domain.

User Interface: We have developed a Web based interface that allows the journalists to access to all the system functionalities.

The components of the NEWS system interact by means of Web Services. The combination of Semantic Web technologies with Web Services results in a quite flexible and open architecture easy to extend, to integrate into the legacy agencies' workflows and with good interoperability properties with other systems. In this paper, we will describe all the NEWS components. A first version of them is currently being tested in real conditions in EFE news agency. The rest of this paper is organized as follows: section 2 describes the annotation component, section 3 describes the HDDDB component and section 4 describes the NEWS GUI. Finally, sections 5, with references on some related work, and 6, with concluding remarks, close this paper.

2 Annotation Engine

NEWS realized the automated annotation of news items using NLP engines for several languages. While originally focussed on text this can also support multimedia news item processing. The next two subsections provide a brief technical

introduction to the annotation engines and to some experiments that we have performed in the context of NEWS with multimedia news item processing.

2.1 Natural Language Processing Engine

In the context of the NEWS project we are using the natural language processing engine of Ontology Ltd., which provides news item categorization, abstract generation and named entity recognition (persons, organizations, locations). This engine categorizes into the International Press Telecommunication Council [3] (IPTC) categorization system known as Subject Code NewsCodes, (a.k.a. Subject Reference System or SRS) which contains more than 1300 categories.

The engine is based on a hybrid approach to natural language processing, which combines linguistic techniques, based on patterns and linguistic rules, with statistical techniques, used to automatically identify and evolve the linguistic rules. There are a diversity of features that the statistical model of Ontology Ltd. learns and later uses to process texts:

Word stems: The Ontology Ltd. annotation engine performs the linguistic tasks of stemming and part-of-speech tagging (a.k.a morphological normalization).

Complex features: These are pairs, triples and higher-order tuples of related words, e.g. the verb *buy* with the noun *company* serving as its object. Such pair of words conveys a more specific notion than the isolated terms *buy* and *company* and should be treated as one complex feature. The detection of this and other syntactic role dependencies in a large variety of phrase structures is performed using linguistic processing.

Concept classes: Words and expressions that in a specific context play similar roles are grouped into one class and are analyzed statistically as a collective feature (in addition to their treatment as individual features). In the context of *Mergers & Acquisitions*, for example, one may want to consider the shared meaning of the verbs *buy*, *sell*, and *purchase* and the noun *acquisition*. Concept classes can be organized in a hierarchy, either as a tree or as a more general graph, allowing a class to have more than one conceptual parent.

All types of features coexist in the categorization and extraction engines of Ontology Ltd. and interact with each other. In the NEWS project, Ontology's engines process news in Spanish, English and Italian languages.

2.2 Dealing with Multimedia News Items

News agencies not only produce textual news items, but also multimedia items such as photographs, audio and video. In the context of the NEWS project, we have analyzed how these multimedia items can be processed by the natural language processing engine in order to generate automatic annotations about them.

Photographs in a news agency have normally attached a textual description, and some management metadata, such as the date, categories, location or author. The NEWS system can analyze the textual description and annotate persons, locations and organizations related to the photograph. We are evaluating the integration of the NEWS system with the photograph archive of the Spanish news agency EFE, called *Fototeca*. The archive contains now over 2 million photographs and about 1500 new ones are added daily into it.

Regarding audio and video news items, our approach consists in using a speech-to-text tool to transcribe the content of the audio track. The resulting text can be processed by the natural language processing engine. The state-of-the-art speech-to-text tools provide high-quality transcriptions and support many languages. In the environment of NEWS, however, there are some drawbacks: major speech-to-text tools require the system to be previously trained by the speaker, but training for every journalist is not possible; audio tracks of news items are normally recorded in noisy environments with portable equipment, and transferred to the agencies by low-quality channels (e.g. conventional or cellular telephone lines); punctuation marks and capital letters normally need to be dictated in the speech, which is impossible in recorded news items.

We have performed two experiments in order to evaluate the impact of the above-mentioned drawbacks in the quality of the annotations, using a commercial speech-to-text tool: *Dragon Naturally Speaking*. In the first experiment, we trained *Dragon* for one male user and one female user. We applied it to five different audio news items provided by the EFE news agency. When the gender of training and the one of the analysed speaker match we obtained a rate of 70%–80% of correctly transcribed words. Gender mismatch however lead to very poor results.

In a second experiment, we evaluated the impact of the loss of punctuation marks. Our experiment was run over a collection of 171 news items. Each news item was processed twice: once the original version and once without punctuation marks. Comparing annotations of the two versions of each news item, we found that the average error rate is about 20%, considering that an error is an annotation in the original item nor present in the modified one, or vice versa.

Although not perfect, experiments show that our approach produces medium-quality annotations without human intervention, and is a feasible alternative to a high-cost manual annotation system.

3 Heuristic and Deductive Database

The Heuristic and Deductive Database component has three main functions in the NEWS system architecture:

- It acts as a Deductive Database (DDB) which stores the news items and their metadata. This repository can be queried by users (pull distribution mode), or used to implement subscription-based push services. It consist basically of three elements:

- A relational database, which stores the information in the NEWS Ontology and the metadata of the news items.
 - A text indexing engine, Lucene [4], used to allow keyword-based queries over the textual contents of news items.
 - An inference engine which allows basic reasoning (transitive closure computation) with the NEWS Ontology.
- It uses some heuristics in order to perform context-aware instance identification (so the H in HDDB), mapping entities found by the natural language processing tool to instances in the NEWS Ontology.
 - It uses the information inside the NEWS Ontology and the Lucene index in providing basic event recognition features.

The next three subsections will describe in more detail some of the key components of the HDDB: the NEWS Ontology, the instance identification algorithm and the event recognition feature.

3.1 The NEWS Ontology

As stated in section 1, the main objective of the NEWS project is to apply Semantic Web technologies in the journalism domain. Applying Semantic Web technologies to a certain domain of interest requires the definition of ontologies to model such domain. In the context of the NEWS project, we have developed the NEWS Ontology. It is a lightweight RDFS [5] ontology which is composed of three modules:

Categorization Module: Provides the vocabulary for automatic news categorization. It is based on IPTC's Subject Code NewsCodes. The Subject Codes set is a three level hierarchy that consists of Subjects, Subject Matters and Subject Details. The top level contains 17 different Subjects. The full set has about 1300 categories. In the categorization taxonomy module, concrete categories from NewsCodes are defined as RDFS classes.

Management Metadata Module: Taking as basis DC [6], PRISM [7], the management metadata included in news representation standards as NITF [8] or NewsML [9] and the management metadata currently used by news agencies, we have developed a vocabulary to be used in annotating news items with management metadata. It covers, among others, topics like authorship information, news item priority and news item media type.

Content Annotation Module: Provides the basic vocabulary for news content annotation. As in principle almost anything in the world can appear in a piece of news, we have decided to rely on a generic top-level ontology. It is inspired in SUMO/MILO [10] and consist of more than 200 classes and about 100 properties. The module was also populated with instances taken from different information sources like ISO country codes standard ISO 3166:1997-1, CIA WorldFact Book [11], NASDAQ companies codes, SUMO/MILO instances, Web sources as Wikipedia [12], etc. As a result the current version contains more than 11,000 instances of different classes: countries, cities, companies, persons, etc. In order to deal with multilingualism,

we have added to the different components in the ontology labels in Italian, Spanish and English. To relate the labels with the ontology elements, we are using Simple Knowledge Organization System, SKOS [13] properties (preferred label, alternative label, description).

3.2 Context-Aware Instance Identification

As we have seen, the NLP engines are able to extract basic entities from text. But in order to achieve fine-grained retrieval results it is not enough to figure out, that the extracted text string *Bush* represents a person, we need to know who is that person by mapping the entity with an instance in the NEWS Ontology.

In order to deal with this problem, the NEWS consortium has developed the IdentityRank algorithm based on PageRank [14]. Basically this algorithm exploits all the information provided by the natural language processing engine (categories, entities) and the news item timestamp as context for entity disambiguation. It is based on two principles:

Semantic coherence: Instances typically occur in news items of certain SRS categories, e.g. president *Bush* in news items of politics category. Also the occurrence of a certain instance gives information about the occurrence of other instances. For example, the spanish F1 driver *Fernando Alonso* usually appears in news items where the F1 Team *Renault* is also mentioned.

News trends: Important events typically are described with several news items covering a certain period of time. For instance when the former Pope died, news items describing such event were composed during several days, most of them including instances as *Vatican* or *John Paul II*.

The process to run the algorithm starts by defining a set of candidate instances for each entity detected in the news item by the natural language processing engine. This can be done by matching the entity text with the labels (in the appropriate language) of the NEWS Ontology's instances. The entity type should also be matched with the instance class to define a candidate instance, so for example we do not consider an instance of class *Region* to be a candidate for an entity of type person.

Once we have all candidate instances of all the entities, we define a semantic network with all these instances. In such semantic network nodes represent candidate instances for the entities in a news item and arcs between two nodes appear when the two instances have cooccured in the past in at least one news item. The category and the timestamp information are taken into account in giving initial weight to the nodes in the semantic network, so if an instance has recently appeared in news items, its node is given more weight. If the instance usually appears in news items of the same SRS category/ies as the news item being considered, its node is also given more weight.

Once the semantic network is defined, we apply the IdentityRank to it. In contrast to classical PageRank, the weight of a node or instance is not divided uniformly between all adjacent nodes, but we give weights to arcs. If two instances typically cooccur, a bigger weight is given to the arc which connects

these instances. In any case, both algorithms deliver a ranking of the nodes in the network. By looking at such ranking, we can find in our case the most relevant instances which can be mapped to the entities in the news item.

We have carried out a basic experiment with 30 news items from EFE news agency where the entity *Georgia* of type location appears with an ambiguous meaning, representing either the US state or the Asian country. The results of the experiment indicate that the algorithm produces a correct mapping entity-instance in about the 80% of the situations, but of course more extensive tests need to be performed.

3.3 Event Recognition

News agencies produce news items that describe events, so automatic event recognition is an attractive feature for them. In order to deal with this requirement, in NEWS we had to deal with two complementary aspects: defining what an event type is at the NEWS Ontology and implementing an event type recognition facility.

In NEWS, an event type definition includes a set of SRS categories, which are expected to be associated with the event, and a set of language dependant patterns manually defined by the ontology maintainer. These patterns have an SPO (Subject-Predicate-Object) structure, where subject and object are defined as one of the following three entity types: person, organization or location, and the predicate can include a verb and/or some free text. The P part of the SPO pattern is compulsory, but S and O are optional. Examples of valid patterns could be the following:

S-P-O ORG - (buy|bought|has_bought) - ORG

S-P PER - (say|said|has_said)

P-O (visit|visited|has_visited) - LOC

The event recognition software is based on the Lucene *span query* facility, so uses the information in the Lucene index at the HDDB. Basically a *span* provides information about where a match of a certain token (piece of text, usually a word) has taken place in a document. That position information can be used later in queries. For the purposes of this paper we are mainly interested in the following kinds of Lucene span queries¹:

SpanTermQuery: Matches spans containing a term. Basically allows to find a certain term or token in a document and its position/s.

SpanNearQuery: Matches spans which are near one another. Developers could specify the *slop*, the maximum number of intervening unmatched positions, as well as whether matches are required to be in-order. For instance, a *span near* query could be built by combining two *span term* queries and providing the *slop*, the maximum number of positions between the results of such queries.

¹ See the Lucene 1.9.1 package `org.apache.lucene.search.spans`.

SpanOrQuery: Matches the union of its clauses. It is used to combine (with OR logical operator) other span queries.

So, the patterns that we have introduced previously are converted inside the Hddb to rules, each of which consists of a set of span queries to Lucene of the types described above. For instance given the pattern:

ORG buys|bought|has_bought|will_buy ORG

and assuming that the natural language processing engine has found the following entities of type organization: *Intel Corporation* and *Microsoft* inside a news item, the following query would be generated to check the news item for occurrences of the pattern and the corresponding event type:

```
SpanNear(
  SpanNear(SpanTerm(Intel), SpanTerm(Corporation), SLOP=1),
  SpanOr(
    SpanTerm(buy),
    SpanTerm(bought),
    SpanNear(SpanTerm(has), SpanTerm(bought), SLOP=1),
    SpanNear(SpanTerm(will), SpanTerm(buy), SLOP=1)
  ),
  SpanTerm(Microsoft),
  SLOP=15
)
```

If the news item matches the query and is an instance of one of the SRS categories associated to the event (or of one of their subcategories) then an association between the event type and the news item is stored in the relational database of the Hddb.

Note that the mechanism that we have described is based on text queries, so it is language dependent. We make it compatible with our requirement of multilinguality by providing versions in spanish, english and italian for each pattern. In order to make the process of pattern definition simpler to the ontology maintainer, we have implemented a Web Service which receives as input a pattern with a verb in infinitive form and generates variants for the pattern and translations to the different languages. For instance, given the pattern in italian: *bambini vb:giocare calcio* (note that the infinitive in the initial pattern is marked with *vb:*) the suggestion service provides the following alternatives:

```
en: children (play|played|has_played) soccer
es: niños (juega|jugó|ha_jugado) fútbol
it: bambini (gioca|ha_giocato) calcio
```

As automatic translation is used in pattern generation, the ontology maintainer should carefully check the results of the process.

4 Graphical User Interface

In order to allow the journalists to access all the NEWS system functionalities, we have developed a Web based graphical user interface accessible via conventional

browser. This interface is implemented using Java Server Pages which are served by an Apache Tomcat [15] engine.

The first step that the journalist should perform in order to access the system is authentication. After that, (s)he will have some tabs organized according to the functionalities offered by the system. These tabs are:

Latest News: Push news item retrieval based on user profile.

Browse Knowledge Base: Pull news item retrieval by browsing the NEWS knowledge base.

Query: Pull news item retrieval by querying.

Create NewsItem: News item creation.

Edit Profile: Personalization and user profile definition.

4.1 Latest News

This is the tab being shown by default when the user logs in into the system. Basically its role is to show the latest news items stored in the Hddb filtered according to the information on the user profile (categories, languages, date ranges). These news items are sorted by timestamp, so the first being shown to the user is the most recent one. The information in this tab is refreshed periodically with a period defined by the user in his/her profile.

4.2 Browse Knowledge Base

The NEWS GUI allows different possibilities for browsing the knowledge stored into the Hddb. These are:

Browse by Categories: Allows the user to find the news items that belong to a set of categories of the IPTC categorization system and that have occurred in a specific date range. The user can select one or more categories from any of the levels and (s)he will get the news items that belong to any of those levels or to one of their sublevels, and that occurred in the specified date range.

Browse by Instances: Allows the user to get news items that contain a concrete instance. The tree with the classes of the NEWS Ontology content module is displayed. The user can browse this tree to select the class that contains the instance (s)he looks for. After clicking on the name of the class, (s)he has to click on a "Find Instances" button. This will retrieve all the instances that appear in any news item stored in the database in the selected date range. Finally, the user has to select one of the instances in order to get the news items that contain this instance in the selected date range.

Browse by Properties: Performs a function similar to that of the previous one: find news items that contain instances. However, in this case, these instances are not selected directly, but through the use of properties. First of all the user selects the class of the instance (s)he looks for (for example the class *Human*). When clicking on that class a list of the properties which have

the class as domain or range (for instance *Human works_at Organization*) appears. By selecting a certain property (like *works_at*) a set of instances are shown to the user. If the initially selected class is the domain of the property, the instances are those in the range (*Organizations* in our example) and vice versa. By clicking on an instance (for example *Ferrari*) the user is looking for news items which mention the instance/s of the initially selected class which are related with the selected instance by the selected property (that is, news items talking about *Humans* that *work_at Ferrari* in our example).

Browse by Events: This tab (see figure 1) allows users to get news items related to a concrete event. Events (see 3.3) are related to the IPTC categorization system so the tree with these categories is shown in order to ease the task of finding them. Once the category is selected, a query is sent to the HDDB to get the events that are related to the category. Finally, the user has to choose one of the available events before sending the query and retrieving news items.

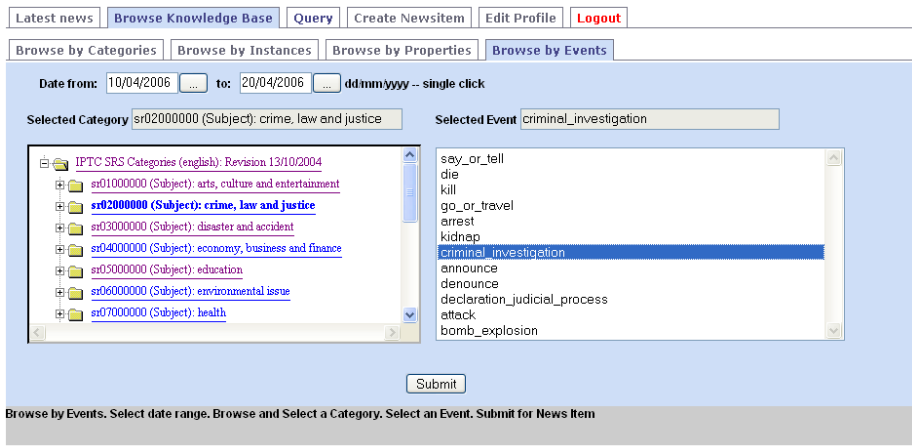


Fig. 1. NEWS Browse by Events Window

4.3 Query

This tab allows the user to introduce queries in order to retrieve news items that satisfy certain criteria. Two kinds of queries can be performed: keyword-based basic queries and advanced queries, where additional restrictions on metadata information can also be specified. The process of performing a basic query is the following:

1. The user enters the terms of the query in the text field.
2. The system tries to match the terms of the query with known instances or entities in the knowledge base in order to perform a semantic search. If there are several possibilities, the user is requested to disambiguate the query.

3. The system looks in the knowledge base for news items that match the entities/instances selected in the previous step. If there are not selected entities nor instances, a full text search is performed.
4. The user can click on a result in order to retrieve the full text of the news item and its metadata, including the entities and instances it is related to.

The advanced query interface permits restricting these results by filtering them by SRS category, news agency, etc.

4.4 Create NewsItem

This tab guides the journalist through the news item creation process, which includes:

1. Manual typing of the content of the news item and some management meta-data like priority or edition location.
2. Automatic categorization and annotation of the news item. The results are shown to the user who can validate both the categories and the entities recognized by the annotation engine. During this step the instance identification process takes also place using the information provided by the NLP engine and a Web service at the HDDB to run the IdentityRank algorithm. The results of this process can also be validated by the journalist. The GUI also allows the manual addition of new categories, entities and instances to the news item.
3. Automatic recognition of event types using a Web service at the HDDB and event type validation (and/or addition) by the journalist.

Finally, the news item is sent to the HDDB where it is stored.

4.5 Edit Profile

One of the functionalities required by the news agencies was the possibility of adapting the system to the user. In order to achieve this task the GUI provides the Edit Profile tab. In this tab the user can modify, among others, the following properties:

- Language of the user interface.
- Refresh time of the Latest News tab.
- Preferred categories, language/s and date range for news item filtering in Latest News tab.

5 State of the Art

The news domain has a number of features that make it interesting for running experiences using Semantic Web technologies in real business: data heterogeneity, huge amounts of information to manage, multilinguality, etc. Taking this into account it is not strange to find in the state of the art several projects related with the topic of applying Semantic Web technologies to the journalism domain, like for instance:

- NAMIC.** The News Agencies Multilingual Information Categorisation project [16] had as main objective to develop and bring to marketable stage advanced NLP technologies for multilingual news customisation and broadcasting throughout distributed services. Though dealing with multilingual and categorization issues in the context of professional journalism, as NEWS, the application of technologies as ontology-based reasoning or a Web Service based architecture were not considered.
- KIM.** The KIM Platform [17] provides a Knowledge and Information Management (KIM) infrastructure for automatic semantic annotation, indexing, and retrieval of unstructured and semi-structured content. Though the KIM platform can be integrated in different contexts, it has not been designed taking news agencies' specific requirements, as compatibility with journalism standards like SRS, into account.
- PENG.** The Personalized News Content Programming project [18,?] aims at defining a news content composition and programming environment that provides news professionals with a flexible tool for a user customizable filtering, retrieval and composition of news. Though dealing with the news domain, this is done from a general perspective so again some specific requirements of the news agencies like multilinguism are not directly covered.
- MESH.** The Multimedia Semantic Syndication for Enhanced News Services project [20] will create multimedia content brokers acting on behalf of users to acquire, process, create and present multimedia information personalized (to user) and adapted (to usage environment). It is an EU funded project that has just started (march 2006) so at the moment there are not available results that can be compared with NEWS achievements.

6 Conclusions and Lessons Learned

In this paper we have described the main achievements of the NEWS project, which has as main objective to bring the Semantic Web technologies into the news domain. Among these achievements we can cite:

1. The development of an ontology for the domain of interest.
2. The design and implementation of annotation mechanisms (categorization, entity extraction and event recognition) in the languages of interest (English, Spanish and Italian).
3. The development of algorithms which allow to match entities detected by the annotation engine to instances in an ontology.
4. The development of intelligent components which exploit the annotations and the information in the NEWS Ontology to perform semantic information retrieval dealing with the multilinguism requirement.
5. The design and implementation of proper user interfaces that allow the journalists to access all the system functionalities in a personalized manner.

A first prototype of the NEWS system has been recently deployed at EFE. We refer the interested reader to [21] for more information on that. Another

prototype is available for public access at [22]. As main lessons learned during the project lifecycle we can cite the following:

Integration into existent workflow: NEWS components were required to be easily integrable and interoperable with legacy tools and workflows. In order to do so, we have developed our components as Web services, which provide a modular and flexible solution. For instance, in principle it is possible that the agencies use only some components, or replace in the future ours with others performing similar operations.

Response time: It is crucial in the news domain, where freshness information is a very important concern. In the context of NEWS this requirement of news agencies had one important consequence: reduce reasoning process, which is complex and time consuming. In our case, we have reduced reasoning to query expansion over the NEWS Ontology.

Scalability: Our applications should be able to handle thousands of new news items each day, and to manage repositories containing millions of items. The consequences in NEWS were clear:

- Use as much as possible well-known scalable technologies as relational databases and classical text indexing engines.
- Avoid reasoning to do things which can be easily implemented and performed by classical procedural mechanisms.
- Use offline mechanisms to perform complex operations if possible. For instance, the training process of Ontology Ltd. engine is performed previously to deployment in news agency.

Human Interface: If, as it is the case of NEWS tools, a non technician human user is going to interact with our systems, the design of the human interface is a crucial issue. Multilingual issues, usability, reliability and completeness (it has sufficient options to access all the available functionalities) are all factors to be taken into account. In that sense our experience when deploying the system in news agencies was that classical tree-based ontology browsing mechanisms (that for instance we provide at the GUI to browse the SRS tree and the NEWS Ontology content module tree) were not well accepted by journalists because for them it is sometimes difficult to know where to find the class that they need. As a possible workaround to this problem we provided keyword based navigation to allow to the journalists to look for a specific class by name, but using personalized views of the ontologies is a possibility to be explored in the future.

Acknowledgements

This work has been partially funded by the European Comission under contract FP6-001906 in the framework of the Information Society Technologies (IST) programme.

References

1. NEWS Home: <http://www.news-project.com>
2. Fernández-García, N.; Sánchez-Fernández, L.; Building an Ontology for NEWS Applications. Poster in the 3rd International Semantic Web Conference, ISWC04.
3. IPTC Home: <http://www.iptc.org/>
4. Apache Lucene Home: <http://lucene.apache.org/>
5. RDF Vocabulary Description Language 1.0: RDF Schema.
<http://www.w3.org/TR/rdf-schema/>
6. Dublin Core Metadata Initiative (DCMI). <http://dublincore.org/>
7. PRISM: Publishing Requirements for Industry Standard Metadata.
<http://www.prismstandard.org/>
8. NITF: News Industry Text Format. <http://www.nitf.org/>
9. IPTC NewsML Web. <http://www.newsml.org/>
10. Niles, I. and Pease, A.; Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Ogunquit, Maine, October 17-19, 2001.
11. CIA WorldFact Book. <http://www.cia.gov/cia/publications/factbook/>
12. Wikipedia Home: <http://wikipedia.org/>
13. Simple Knowledge Organisation System, SKOS.
<http://www.w3.org/2004/02/skos/>
14. Page, L.; Brin, S.; Motwani, R and Winograd, T.; The PageRank Citation Ranking: Bringing Order to the Web. <http://dbpubs.stanford.edu/pub/1999-66>.
15. Apache Tomcat Home: <http://tomcat.apache.org/>
16. NAMIC: News Agencies Multilingual Information Categorisation.
<http://www.dcs.shef.ac.uk/nlp/namic/>
17. Kiryakov, A.; Popov, B.; Ognyanoff, D.; Manov, D.; Kirilov, A. and Goranov, M.; Semantic Annotation, Indexing and Retrieval. In Proceedings of the 2nd International Semantic Web Conference, ISWC 2003, LNCS 2870, pp 485-499.
18. PENG: Personalized News Content Programming. <http://www.peng-project.org/>
19. Pasi, G. and Villa, R.; Personalized News Content Programming (PENG): A System Architecture. In proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05).
20. MESH: Multimedia Semantic Syndication for Enhanced News Services.
http://cordis.europa.eu/ist/kct/fp6_mesh.htm
21. Sánchez-Fernández, L.; Fernández-García, N.; Bernardi, A.; Zapf, L.; Peñas, A. and Fuentes, M.; An experience with Semantic Web technologies in the news domain. In Proceedings of the ISWC 2005 Workshop on Semantic Web Case Studies and Best Practices for eBusiness, SWCASE05, CEUR-WS, Vol. 155.
22. NEWS Web Demo. <http://corelli.gast.it.uc3m.es:8081/NEWS/> (login: guest, password: guest)