

An experience with Semantic Web technologies in the news domain

Luis Sánchez-Fernández¹, Norberto Fernández-García¹, Ansgar Bernardi², Lars Zapf², Anselmo Peñas³, Manuel Fuentes⁴

¹ Carlos III University of Madrid, Leganés (Madrid), Spain
<luis,berto@it.uc3m.es>

² German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern,
Germany
<bernardi,zapf@dfki.uni-kl.de>

³ UNED, Madrid, Spain
<anselmo@lsi.uned.es>

⁴ EFE S.A., Madrid, Spain
<mfuentes@agenciaefe.com>

Abstract. The news domain is an interesting area to experiment with Semantic Web technologies, because there is the need to manage huge amounts of content. In this paper we present an experience of deploying Semantic Web technologies to the Spanish news agency EFE. These technologies include semantic annotation and intelligent information retrieval, and are developed as part of the NEWS European project.

1 Introduction

The news domain has a number of features that make it interesting for making experiences in using Semantic Web technologies in real business. News agencies produce huge amounts of content in the form of news items describing an event. Most of this content is text, but they also produce multimedia content in different human languages. The success of news agencies as companies relies on their ability to manage all this heterogeneous information in an efficient manner.

As partners of the project NEWS (News Engine Web Services) [1], we work on the usage of Semantic Web technologies to help news agencies to deal with this challenge. NEWS is a research and development project funded by the European Commission under contract FP6 001906 in the framework of the Information Society Technologies (IST) programme. The project consortium is composed of two news agencies, EFE¹ and ANSA², the DFKI³ research institute, Ontology Ltd.⁴ company, and Carlos III University of Madrid⁵. Its main objective is to

¹ <http://www.efe.es>

² <http://www.ansa.it>

³ <http://www.dfki.de/web/>

⁴ <http://www.ontology-ltd.com>

⁵ <http://www.uc3m.es>

develop semantic based tools and systems to be deployed in news agencies in order to improve their productiveness and revenues.

In this paper we present an experience of deploying Semantic Web technologies in one of the industrial partners of the NEWS project: the Spanish news agency EFE. The technology to be described is currently being developed in the framework of the project, and, as will be shown, includes semantic annotation and intelligent information retrieval tools.

The rest of the paper is organized as follows: section 2 describes EFE news agency and briefly introduces the main steps in the news item management process. It also describes the possible benefits of using Semantic Web technologies in a news agency. Section 3 introduces the NEWS project, describing the most important components which are being developed in it. Section 4 describes how the NEWS components are deployed in EFE news agency. Section 5 introduces, with discussion purposes, our experiences in the development and deploying process. Finally, section 6 with concluding remarks finalizes this paper.

2 Scenario: The EFE News Agency

EFE Agency (Agencia efe) is a company focused on the international market, specially Spain in Europe, South America, and the Spanish speaker media from the United States. EFE has clearly a complete multimedia vocation; nowadays, EFE works with all the formats, text, photo, audio and video.

EFE has 25 offices in Spain, 35 abroad and a wide net of small branches and correspondents in 150 cities of 100 countries. EFE staff is composed of 1000 people and several thousand external collaborators. The daily production is more than 3000 text news in Spanish, Catalan, English, Brazilian and Arabic language, 1000 pictures, and an important variety of video and audio news services. The main commercial data bases are the so called EFEdata and Fototeca. The first of them stores more than 9 million of text news with document from 1988 and the second contains 1.400.000 pictures from 1998, ten years later.

The steps in the *lifecycle* management process of an EFE news item are:

Generation The news item contents are edited using the news agency edition tools, which in the case of EFE are most of them proprietary. Apart from the contents of the news item itself, these edition tools allow to the journalists to annotate the news items by filling in some forms. This manual annotation process includes the addition of information as the author, creation time and date, keywords, and creation location. Multimedia news items are also annotated with a textual description of its contents. Another metadatum which is also added to every news item, is the category or categories of such news item in a certain categorization system, like ANPA-EFE, the internal EFE system. Typically this category reflects an issue which is covered by the news item, so for instance ANPA-EFE has categories for sports news items, economy news items, etc.

Once the news item content and annotations are provided by the journalist, a new news item is generated. News items are represented as documents in

a certain format. The formats used in the recent past were plain text based, like ANPA [2], and binary ones, like IIM [3], but nowadays EFE is starting to use XML-based formats like NITF [4] and NewsML [5], both standardized by IPTC, International Press Telecommunication Council [6].

Storage After news items are generated, they are stored in the data warehouse infrastructure of the news agency, which includes different kinds of textual, relational and documental databases. Of course redundancy storage and backup techniques are applied, in order to guarantee the survival of such valuable information to any unexpected situation.

Distribution Basically EFE news items are delivered to customers in two ways: by satellite broadcasting, using a VSAT (Very Small Aperture Terminal) network with world wide coverage, and by means of special push/pull services based on terrestrial IP networks. In push model, clients subscribe themselves to a service, and the news items which are relevant to that service are sent to them. In order to decide if a news item is relevant to a service, the metadata of such news item can be used. For instance, it is possible to have services specialized in a certain sector (economy, sports, politics, etc) and using the category/ies of the news item help to decide if the news item is relevant to a certain service. In the pull model, the users can query EFE databases looking for news items with certain properties, and buy only those that are interesting to them. So, for this second kind of services, information retrieval tools are crucial in order to provide to customers the news items which best match with their queries/interests. At this moment, EFE is using two different information retrieval tools: PLS (from the, now disappeared, Personal Library Software company) and Autonomy [7].

Analyzing more in detail the scenario that we have depicted, it seems to us that in some points the current EFE workflow could be improved, for instance:

Categorization At this moment the categorization is done by hand and using basic categorization systems, like ANPA-EFE. This task could be improved by using an automatic categorization tool, which also allows the usage of more complex categorization systems like the one provided by the IPTC: the Subject Codes NewsCodes [8].

Content annotation The annotations currently added to the news items are mostly for management purposes. The contents of the news items are not annotated, so for example, the basic entities (organizations, persons, places, etc) which are mentioned inside the news item are not tagged. These content annotations could be used in fine-grained news item selection, allowing the development of more advanced push services, like, for instance, a service providing news items talking about a certain person or organization.

Multilinguality As we have seen, EFE is currently using two different information retrieval tools: PLS and Autonomy. One of the main drawbacks of these tools is that they work with text, which is language dependant. As EFE produces contents in several languages, it could be interesting to have some kind of multilingual information retrieval tool. For example, this tool

would allow to customers looking for news items talking about *Rome* to retrieve news items in different languages, even if they do not contain the text string *Rome*, but for instance *Roma*⁶.

3 The NEWS Project

The NEWS 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 the NEWS project covers mainly three topics:

Ontology development Using Semantic Web standards to define ontologies for the news industry. We have developed the NEWS Ontology [9], which covers the main concepts required in the news domain. It is a lightweight RDFS [10] ontology and provides the basic classes, properties and instances for news item categorization and content annotation. In its design we have taken into account the standards from the journalism world, which have been used as sources in the initial knowledge capture process. Another point which we have also considered is multilinguality: the concepts included in the NEWS ontology have associated labels and descriptions in several languages (more specifically, Spanish, Italian and English). These labels and descriptions are useful, for instance, in order to implement semantic search facilities, which require that users disambiguate their queries.

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 developed by Ontology Ltd. This engine uses an hybrid approach to natural language processing, based on combination of morphological and syntactical analysis with statistical tools. It provides automatic news item categorization using the classes of the NEWS Ontology, which are basically the ones defined by the Subject Codes NewsCodes. The power of the NEWS classification engine is demonstrated on a live demo [11] using news items harvested from BBC and Times web sites among others. Another facility provided by the annotation component is named entity recognition. The basic entities currently covered are those which are considered most relevant by news agencies: persons, organizations and locations. Both named entity recognition and categorization can also be used with multimedia news items, which have attached a text description processable by the Ontology Ltd. engine.

Intelligent Information Retrieval Developing news intelligent components, with multilingual and multimedia capabilities, which use semantic annotations and ontologies to allow the development of intelligent news services. In the context of NEWS we have developed a Heuristic and Deductive Database (HDDB) component which has two main tasks:

⁶ Spanish word for Rome.

Deductive Database The HDDB acts as a repository of news items, which can be queried by users (pull mode), or used to implement push services. Basically this repository consist of three components: a relational database, which stores the instances of the ontology and the metadata of the news items, a text indexing engine, Lucene [12], used to allow keyword-based queries, and an inference engine, which provides the deductive part of the system. This inference engine relies on the NEWS Ontology for semantic query expansion, and is based on the TRIPLE system [13] which uses XSB Prolog [14] as reasoner. The HDDB component integrates all these elements in order to allow the combination of semantic queries with full-text queries, providing flexible and reasonably powerful query capabilities.

Entity Identification The annotation component provides entity recognition facilities, but does not identify the entities. For instance, it says that the text string "Bush" represents a person, but does not say who is that person. The entity identification task associates NEWS Ontology instances with entities recognized in the news item by the annotation component. In order to do so, we are using some heuristics (so, the H in HDDB) which try to exploit the context information in the news item (category, other entities, information in the ontology, etc). This information, the relation between a news item and a NEWS Ontology instance, is later used in order to implement semantic information retrieval.

In the next section it will be shown how all these components are integrated with one another and deployed inside the already existent EFE workflow.

4 NEWS Workflow at EFE

The different components being developed in the NEWS project act together within an application workflow. This workflow has been designed having in mind the requisites of the EFE and ANSA news agencies. As we are currently testing it at EFE, we will center our exposition in the EFE scenario, but a similar workflow is also applicable into the ANSA one.

The workflow design has taken into account that the journalists in the news agencies want to have control over all the content production process. This leads to a semiautomatic solution, where the journalist can validate the result obtained in different processing stages of the news items.

Figure 1 shows the proposed NEWS workflow. The NEWS workflow starts with a journalist creating a news item using agency's edition tool. In our case, the result of this process is a NITF document with the news item contents plus some management metadata like the date of creation, the author, etc.

The news item is then submitted to the semantic annotation component, where categorization and named entity extraction takes place. As a result, the NITF document is sent back enriched with one or several categories plus a number of entities. For each entity it is provided the tagged text, its position in the news item content plus the entity type (person, place or organization).

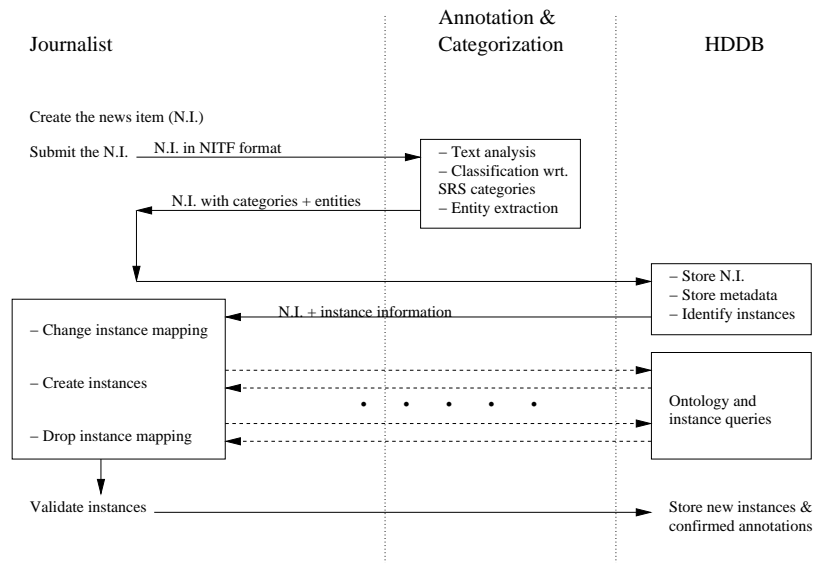


Fig. 1. NEWS Workflow

The NITF document is then sent to the HDDB. There, the news item and the metadata (management metadata, categories and entities) are stored. Also in this step, the instance identification heuristic is executed, and candidate mappings between entities and instances in the ontology are generated.

The candidate mappings between the entities extracted by the annotation system and the instances in the ontology are then sent to the journalist. The journalist can then, by means of a specific GUI, take the following actions:

- Change a mapping, selecting another instance.
- Create a new instance in the ontology and map an entity to the new instance.
- Drop a mapping (in that case only the entity is stored in the HDDB).

At the end, the final step is done when the journalist sends to the HDDB the annotations of instances that occur in the news item plus (possibly) a number of new instances to be added to the ontology.

5 Discussion: Our experience

A first version of the NEWS prototype has been recently deployed at EFE. This section describes some problems we have found and the way we have dealt with some of them in the context of the NEWS project. These are:

Requirements capture From our point of view, one of the problems in this stage was the gap between industrial partners and research partners. At the

time the industrial partners expressed their requirements, they did not know fully the capabilities of Semantic Web technologies, so it was hard for them to discriminate the achievable requirements from the unrealistic ones.

Integration in existent workflow The components developed in the context of NEWS project 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 As we have said, news agencies are data intensive environments. Our applications should be able to handle thousands of new news items each day, and to manage repositories containing millions of already existent news 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 says that it is a good practice to start by analyzing the existing interfaces and try to mimic them as much as possible, reducing the user learning effort and change resistance. Performing internal evaluations and getting direct feedback from users may also help to obtain a good design.

Quality of processes The news agencies need to offer their customers quality content and services. This means, for instance, that the number of errors allowed in the metadata of a news item should be low. As we have said, this restriction has had as consequence in NEWS context resulted in the development of a workflow where automatic processes are human supervised.

6 Conclusions and Future Lines

In this paper we have shown our initial experiences deploying Semantic Web technologies ((semi)automatic annotation and intelligent information retrieval)

in the news domain. We consider this domain attractive for initial experiences. We expect that the use of Semantic Web technologies for news content management will produce benefits in automatization of processes and search and recovery.

Some lessons learnt have been already mentioned. Among them we would like to stress the following: scalability and smooth deployment. Scalability issues should be seriously taken into account when selecting the representation language used for the ontologies and the reasoning capabilities of inference engines. A second lesson is that the deployment of Semantic Web technologies should be done in a smooth way. In our case, the system is currently running experimentally at EFE without affecting the production system. This is possible because the journalist creates the news items using their proprietary tools and the created news item is entered in the NEWS workflow.

Acknowledgements

This work has been partially funded by the European Commission under contract FP6-001906 in the framework of the Information Society Technologies (IST) programme and by the Spanish Ministry of Education and Science under contracts TSI2004-0042-E.

References

1. NEWS (News Engine Web Services) Home. Available at: <http://www.news-project.com>.
2. ANPA Wire Service Transmission Guidelines. Available at: <http://www.naa.org/technology/standard/89-3msw.pdf>.
3. Information Interchange Model (IIM). Available at: <http://www.iptc.org/IIM/>.
4. NITF: News Industry Text Format. Available at: <http://www.nitf.org/>.
5. IPTC NewsML Web. Available at: <http://www.newsml.org/>.
6. IPTC Web Homepage. Available at: <http://www.iptc.org/>.
7. Autonomy. Available at: <http://www.autonomy.com/>.
8. IPTC NewsCodes. Available at: <http://www.iptc.org/NewsCodes/>.
9. Fernández-García, N.; Sánchez-Fernández, L.; Building an Ontology for NEWS Applications. In Poster Session of the 3rd International Semantic Web Conference, ISWC, 2004.
10. RDF Vocabulary Description Language 1.0: RDF Schema. Available at: <http://www.w3.org/TR/rdf-schema/>.
11. Ontology Ltd. annotation engine demo. Available at: <http://www.ontology-ltd.com/demo>.
12. Apache Lucene. Available at: <http://lucene.apache.org/>.
13. Sintek, M. and Decker, S.; A Query, Inference, and Transformation Language for the Semantic Web. In Proceedings of the 1st International Semantic Web Conference, ISWC, 2002.
14. XSB Home. Available at: <http://xsb.sourceforge.net/>.