Building an Ontology for NEWS* Applications

Norberto Fernández-García, Luis Sánchez-Fernández
Carlos III University of Madrid
{berto,luis}@it.uc3m.es

Abstract

In the NEWS (News Engine Web Services) EU IST project we are developing tools to bring Semantic Web technologies into the professional journalism world. As part of this process, we are currently developing an ontology for NEWS applications. In this poster we describe the previous stages of that design, some of the difficulties that we anticipate, and possible solutions to such problems that we are analysing in a first prototype.

1 Introduction

In the current Information Society, being informed is becoming a basic necessity. Due to professional or personal reasons, users are interested in obtaining information of quality wherever they are. Journals, newspapers, and other communication media are required to provide fresh, easily understandable and relevant information to their clients, while they are at home, in a restaurant or just travelling to their jobs.

But media are not the only important provider-actors in the journalism world. Most media are consumers of news agencies, which produce a big percentage of the news published all over the world. Forced by their customers, these agencies must also fulfil some of these requirements.

As partners of the NEWS [NEWS,2004] project, we believe that the usage of Semantic Web technologies and Web Services may help agencies to achieve this goal. So, we centre our work in developing tools which, using previously cited technologies, help agencies in increasing their productiveness and revenues.

As part of the process of implementing tools to apply Semantic Web technologies to the journalism world, we are currently developing an ontology to be used within these tools. As far as we know, this ontology will be the first specifically designed to operate in professional journalism applications. In this poster we describe the previous stages of that process, some of the difficulties that we anticipate, and possible solutions to such problems that we are analysing.

*NEWS is a research and development project funded by the European Comission under contract FP6 001906 in the framework of the Information Society Technologies (IST) programme. The project partners are the news agencies *Agencia EFE S.A.* (1) and *Agenzia ANSA S.C.R.A.L.* (2), the *Deutsches Forschungszentrum für Künstliche Intelligenz Gmbh* Institute (3), *Textology Ltd.* (4) and *Universidad Carlos III de Madrid* university (5)

2 Usage Scenario

As a first step in our process, we have analysed the intended usage scenario of our tools and how can it affect the ontology design. Some conditions introduced by this scenario are:

Ontology Usage NEWS tools will provide among others:

- **News categorisation**: Consists of classifying news items using a taxonomy. The news item class is used, for example, to decide what clients can be interested in a certain item and send it to them (push model). Classification is currently being done by hand and using basic specific taxonomies[1]. In NEWS we propose to automatize such process (with human supervision of results) and define richer taxonomies, using mappings with the old ones, to achieve backwards compatibility.
- **News annotation**: In the scope of NEWS, this will include, not only the whole item metadata currently added (used for management purposes), but also inline annotation of news contents (helpful for example in fine-grained news item selection by clients -pull model-). These inline annotations will be added automatically (again with human supervision of results).

Our ontology will provide the basic vocabulary for annotations and the taxonomy for news item classification. To perform these functions, high degree of formalisation (axioms) seems unnecessary, so we are thinking on developing a lightweight ontology.

News Representation Standards Compatibility
Journalism standards like NewsML [NewsML,2004], NITF [NITF,2004] or IIM [IIM,2004] are currently in use or are expected to be used in the future by news agencies to represent news items. Our metadata model should be compatible with these standards. We have analysed them looking for information about what kinds of metadata (global, inline) they support and where can

[1] For example, taxonomy of ANSA consist of only 11 classes

we add metadata to news items represented using such standards.

News Metadata Standards Compatibility Some standards like Dublin Core (DC) [DC,2004], Publishing Requirements for Industry Standard Metadata (PRISM) [PRISM,2004] or Subject Reference System (SRS) [SRS,2004], could provide us sets of standardised metadata. Compatibility with these standards is highly desirable because using such metadata, agencies can make their contents more accessible to other applications which are also compatible with them (for example, with edition applications from communication media).

Fast Processing One of the requirements shown in introduction section is freshness: news should be delivered to clients as fast as possible. With this requirement on mind we can discard the usage of complex time-consuming inference processes, which reinforces the idea of developing a lightweight ontology. Another aspect influenced by the processing speed is the usage of manual classification or annotation, which is discouraged. This affects the ontology in the sense that we must include in it only the classes and properties which can be recognised in the news text by the automatic annotation and classification tools.

Range of Contents One of the biggest problems of building an ontology which can be used to annotate the contents of a news item, is that almost everything in the world can appear in a piece of news. It seems we need to model all things in the world, which is far from being an easy task. Possibilities to deal with this problem are the usage of already existent general ontologies like OpenCyc [OpenCyc,2004], the usage of a top level ontology (like for example SUMO [SUMO,2004]), independent of application domain, or even the definition of different ontologies for different domains and association of domain specific ontologies to news categories.

Ontology Standards Compatibility We are interested in being compatible with widely accepted ontology-related standards, like W3C's languages for ontology definition: RDFS [RDFS,2004] and OWL [OWL,2004].

3 Prototype

As a second step in our development process, we have designed a first prototype of our ontology. It is currently represented as a formal description in natural language and consist of three main modules:

News Item categorisation We have developed a basic categorisation taxonomy. It is based on SRS standard, which ensures standard compatibility.

Management Metadata Taking as basis DC, PRISM, the management metadata included in news representation standards as NITF and the management metadata currently used by news agencies, we have developed a model or 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 Metadata We have developed a generic top-level ontology to be used in inline content annotation. It is based in works like SUMO [SUMO,2004] and in the inline metadata components included in NITF specification. It covers topics like time, location, events, persons, organisations, etc.

This design is being tested in a basic usage case with the idea of finding problems and design errors at an early stage. The tests consist in using our ontology to manually categorise and annotate information in real NITF news items provided by news agencies. An example of an RDF document representing the metadata of one of such news items can be found in [RDF example,2004].

4 Conclusions and Future Work

In the NEWS project we are developing tools to bring Semantic Web technologies into the professional journalism world. As part of this process we are currently developing an ontology for NEWS applications. In this poster we have described the usage scenario of such ontology and the restrictions that this scenario imposes to our design. We have also introduced a first prototype of the NEWS ontology. Future work will include the development of more complex usage cases, the extension of the ontology, the refinement of its design and its implementation in RDFS or OWL.

References

[NEWS,2004] *NEWS (News Engine Web Services) Home*. Available at: http://www.news-project.com/

[NewsML,2004] *IPTC NewsML Web*. Available at: http://www.newsml.org/

[NITF,2004] *NITF: News Industry Text Format*. Available at: http://www.nitf.org/

[IIM,2004] *Information Interchange Model (IIM)*. Available at: http://www.iptc.org/IIM/

[DC,2004] *Dublin Core Metadata Initiative (DCMI)*. Available at: http://dublincore.org/

[PRISM,2004] *PRISM: Publishing Requirements for Industry Standard Metadata*. Available at: http://www.prismstandard.org/

[SRS,2004] *Metadata: Subject Reference System and NewsML Topicsets*. Available at: http://www.iptc.org/metadata/

[OpenCyc,2004] *OpenCyc*. Available at: http://www.opencyc.org/

[SUMO,2004] *SUMO Ontology*. Available at: http://ontology.teknowledge.com/

[RDFS,2004] *RDF Vocabulary Description Language 1.0: RDF Schema*. Available at: http://www.w3.org/TR/rdf-schema/

[OWL,2004] *OWL Web Ontology Language Reference*. Available at: http://www.w3.org/TR/owl-ref/

[RDF example,2004] *RDF example of NITF news item representation*. Available at: http://www.it.uc3m.es/news-int/example.rdf