# Bringing The IPTC News Architecture into the Semantic Web

Raphaël Troncy

CWI Amsterdam, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
raphael.troncy@cwi.nl

**Abstract.** For easing the exchange of news, the International Press Telecommunication Council (IPTC) has developed the NewsML Architecture (NAR), an XML-based model that is specialized into a number of languages such as NewsML G2 and EventsML G2. As part of this architecture, specific controlled vocabularies, such as the IPTC News Codes, are used to categorize news items together with other industry-standard thesauri. While news is still mainly in the form of text-based stories, these are often illustrated with graphics, images and videos. Media-specific metadata formats, such as EXIF, DIG35 and XMP, are used to describe the media. The use of different metadata formats in a single production process leads to interoperability problems within the news production chain itself. It also excludes linking to existing web knowledge resources and impedes the construction of uniform end-user interfaces for searching and browsing news content.
In order to allow these different metadata standards to interoperate within a single information environment, we design an OWL ontology for the IPTC News Architecture, linked with other multimedia metadata standards. We convert the IPTC NewsCodes into a SKOS thesaurus and we demonstrate how the news metadata can then be enriched using natural language processing and multimedia analysis and integrated with existing knowledge already formalized on the Semantic Web. We discuss the method we used for developing the ontology and give rationale for our design decisions. We provide guidelines for re-engineering schemas into ontologies and formalize their implicit semantics. In order to demonstrate the appropriateness of our ontology infrastructure, we present an exploratory environment for searching and browsing news items.

## 1 Introduction

Nearly every European citizen reads, watches or listens to the news, at home, while commuting to and from work, at work and even as part of their work. As voting citizens, we need to understand local, national and international politics to allow us to cast our vote. As company employees, we need to understand the state and development of local, national and international economies to enable us to understand our markets. As part of our leisure time, we want to know about our favorite sports teams, the lives of our soap idols or the most recent

books available. Nowadays, this information is online, and hence easily accessible from anywhere.

In existing news workflow processes, news items are typically *i)* produced by news agencies, independent journalists or citizen media, *ii)* consumed and enhanced by newspapers, magazines or broadcasters then finally *iii)* delivered to end users. News items are typically accompanied by a set of metadata and descriptions that facilitate their storage and retrieval. However, much of the metadata is lost because of interoperability problems occurring along the workflow. In addition, at the end user interface, opportunities for making use of the available metadata are often lost. Consequently, users are forced to explore news information in environments that contain large amounts of irrelevant, unreliable and repeated information, with insufficient access to background knowledge.

Our ultimate goal is to create an environment that facilitates end-users in seeing meaningful connections among individual news items (stories, photos, graphics, videos) through underlying knowledge of the descriptions of the items, their relationships and related background knowledge. This requires semantic metadata models to improve metadata interoperability along the entire news production chain. The underlying research problem tackled in this paper covers the two ends of the news workflow spectrum: how to model and represent semantic multimedia metadata along the news workflow and the consequences of this modeling at the user interface.

The contribution of this paper is twofold. On one hand, we report on the modeling of the ontologies for the IPTC family of languages and we convert the IPTC NewsCodes into SKOS thesaurus for demonstrating how the news metadata can be automatically enriched and further integrated with the knowledge already formalized on the Web. We generalize our approach and provide guidelines for re-engineering schemas into ontologies and formalize their implicit semantics. On the other hand, we discuss these modeling decisions with respect to their consequences on the end-user interfaces. We present exploratory interfaces for searching and browsing news that require rich semantic descriptions of the data.

This paper is organized as follows. In the next section, we briefly introduce the main news and multimedia standards used by the media industry. Readers who are already familiar with these formats can skip this section. In Section 3, we discuss the existing methods for engineering ontologies from schemas and porting thesauri to the Semantic Web. We also present the existing attempts for integrating multimedia and news ontologies. We detail in Section 4 the steps for building an ontology-based news infrastructure. We discuss the design decisions and we provide guidelines for re-engineering schemas into ontologies. In order to demonstrate the appropriateness of our ontology infrastructure, we present a semantic search system for multimedia news in the Section 5. Finally, we give our conclusions and outline future work in Section 6.

## 2 News and Multimedia Standards

### 2.1 News Standards

Historically, the International Press Telecommunication Council (IPTC) has developed NITF[1] and NewsML, two XML-based languages for describing the structure and the content of news articles. These languages proved, however, to be inadequate to describe all kind of multimedia news and were often judged too verbose. Recently, IPCT has released the News Architecture framework (NAR[2]) which provides the framework for the second generation of IPTC G2 standards.

NAR is a generic model that defines four main objects (`newsItem`, `packageItem`, `conceptItem` and `knowledgeItem`) and the processing model associated with these structures. Specific languages such as NewsML G2 or EventsML G2 are then built on top of this architecture. For example, the generic `newsItem` is specialized into media objects (textual stories, images or audio clips) in NewsML G2.

Finally, IPTC maintains a number of controlled vocabularies called the IPTC NewsCodes, that are used as values while annotating news items. Among others, the Subject Codes is a thesaurus of 1300 terms used for categorizing the main topics (*subjects*) of each news items.

### 2.2 Multimedia Standards

Although the NAR architecture defines the basic concepts for representing the various media (text, photo, audio, video, graphics), a multitude of other standards are used in the media industry [12].

Pictures taken by a journalist come with EXIF[3] metadata related to the image data structure (height, width, orientation), the capturing information (focal length, exposure time, flash) and the image data characteristics (transfer function, color space transformation). Both Kanzaki[4] and Norm Walsh[5] have proposed an RDFS ontology of EXIF and services for extracting and converting the metadata stored in the header of the images.

These technical metadata are generally completed with other standards aiming at describing the subject matter. DIG35[6] is a specification of the International Imaging Association (I3A). It defines, within an XML Schema, metadata related to image parameters, creation information, content description (who, what, when and where), history and intellectual property rights. In collaboration with Ghent University, we have recently modeled these metadata blocks into

[1] News Industry Text Format: `http://www.nitf.org/`
[2] `http://www.iptc.org/NAR/`
[3] Exchangeable Image File Format:
`http://www.digicamsoft.com/exif22/exif22/html/exif22_1.htm`
[4] `http://www.kanzaki.com/ns/exif`
[5] `http://sourceforge.net/projects/jpegrdf`
[6] `http://www.i3a.org/resources/dig35/`

a DIG35 ontology[7], following the same guidelines detailed in Section 4. XMP[8] provides a native RDF data model and predefined sets of metadata property definitions such as Dublin Core, basic rights and media management schemas for describing still images. IPTC has itself integrated XMP in its Image Metadata specifications[9]. PhotoRDF[10] is also an attempt to standardize a set of categories for personal photo management using Dublin Core and a minimal RDF schema defining 10 terms for the `dc:subject` property.

Video can be decomposed and described using MPEG-7, the *Multimedia Content Description* ISO Standard [15]. This language provides a large and comprehensive set of descriptors including multimedia decomposition descriptors, management metadata properties, audio and visual low-level features and more abstract semantic concepts. The ambiguity and lack of formal semantics of MPEG-7 have been largely pointed out, and several OWL ontologies modeling this standard have been proposed and recently compared [18]. Among them, the Core Ontology for Multimedia Annotation (COMM) proposes to re-engineer completely MPEG-7 using DOLCE as upper ontology and multimedia design patterns [1]. From the broadcast world, the European Broadcaster Union[11] (EBU) has recently adopted the NAR architecture for describing videos, providing some extensions in order to be able to associate metadata to arbitrary parts of videos and to have a vocabulary for rights management.

In conclusion, we end up with an environment that uses numerous languages and formats, often XML-based, that leads to interoperability problems within the news production chain itself and that excludes linking to other vocabularies and existing web knowledge resources. We propose to use Semantic Web languages for leveraging all these standards and ease their integration. This requires a proper ontology infrastructure. Based on the related literature detailed in the next section and our own experience, we discuss the rationale of the design decisions and we formulate guidelines for modeling ontologies from existing schemas.

## 3 Related Work

Many approaches have been reported to build ontologies [11]. For example, Uschold and Grüninger methodology [20] provides the general steps for the whole process of ontology engineering while METHONTOLOGY [6] proposes to build the ontology at the knowledge level using a set of *intermediate representations*. Specific methods focus on the conceptualization of the ontology, that is, how to structure the taxonomy of concepts [4]. These methodologies, however, do not consider the (supposedly easier) case where a schema (UML diagrams, XML

[7] `http://multimedialab.elis.ugent.be/users/chpoppe/Ontologies/index.html`
[8] Adobe's Extensible Metadata Platform: `http://www.adobe.com/products/xmp/`
[9] `http://www.iptc.org/IPTC4XMP/`
[10] `http://www.w3.org/TR/photo-rdf/`
[11] `http://www.ebu.ch`

Schema, thesaurus) formalizing already the domain pre-exists but still needs to be ported to the Semantic Web.

### 3.1 Porting Schemas and Thesauri to the Semantic Web

Semantic Web and object-oriented languages are compared in [14] which further explains how to develop ontology-driven software. The "SKOSification" of thesauri in the cultural heritage domain has lead to a general method for porting thesauri to the Semantic Web [22, 3, 2]. This method advocates four steps (preparation, syntactic conversion, semantic conversion, standardization) and provides a number of guidelines for each step. Our method follows the same recommendations and add more guidelines regarding the modeling of existing UML diagrams in OWL ontologies.

The alignment of the resulting thesaurus with existing semantic web resources is particularly addressed in [17], that leads to the AnnoCultor[12] conversion tool. We have used this tool for converting the IPTC NewsCodes into SKOS thesauri.

### 3.2 NewsML and Multimedia Ontologies

Various attempts for building a news ontology have been reported. NEWS[13] is a completed EU project that aims to combine Semantic Web technologies and web services for improving the news agencies workflow. The project has developed a lightweight RDFS news ontology (in English, Spanish and Italian) based on the IPTC Subject Codes for categorizing the news items and on NITF and NewsML for the metadata management [7, 9]. The Neptuno[14] research project has also modeled a lightweight RDFS news ontology representing a newspaper archive. The ontology is again a mix between news management metadata based on the NewsML standard and on the IPTC Subject Codes aligned with a news agency thesaurus for categorizing the news items [5]. Finally, MESH[15] is an ongoing EU project that focuses on multimedia analysis for enriching automatically news metadata and deliver personalized news summary. A news ontology seems to have been developed but it is not available.

In contrast to these projects, our approach is to decouple the thesauri used in the metadata values from the ontology that describes the management of the news items according to the journalist point of view. This separation of concern provides a more flexible infrastructure where the Subject Codes can be aligned to other thesauri. We expose these aligned thesauri on the Semantic Web, providing dereferencable URIs for every terms. Furthermore, we conform to the latest standard for the news metadata (NAR) and we design the ontology to be linked with other media ontologies.

[12] `http://sourceforge.net/projects/annocultor`
[13] `http://www.news-project.com/`
[14] `http://seweb.ii.uam.es/neptuno/`
[15] `http://www.mesh-ip.eu/`

The XML Semantics Reuse methodology consists in converting automatically XML Schemas into OWL ontologies[16]. This methodology is used in the journalism domain for converting the NewsML and NITF document formats, the IPTC Subject Codes taxonomy and the MPEG-7 multimedia format into OWL/RDF [10]. The resulting ontology, however, fails to capture the intended semantics of these standards that cannot be represented in XML Schema [18]. It recreates the complex nested structures used in the original schema (e.g. the definition of intermediate containers defining the XML Schema types and elements) that should generally not be modeled in the ontology. We advocate, on the contrary, to re-engineer the ontology following some good practices that we detail in the next section.

## 4 Building a Semantic Web Infrastructure for News

As we have described in the Section 2, NAR is a generic model for describing the news items as well as their management, packaging, and the way they are exchanged. Interestingly, this model shares the principles underlying the Semantic Web:

- News items are distributed resources that need to be uniquely identified like the Semantic Web resources;
- News items are described with shared and controlled vocabularies.

NAR is however defined in XML Schema and has thus no formal representation of its intended semantics (e.g. a `NewsItem` can be a `TextNewsItem`, a `PhotoNewsItem` or a `VideoNewsItem`). Extension to other standards is cumbersome since it is hard to state the equivalence between two XML elements.

By modeling a NAR ontology, we do expect the following benefits:

- Better control of NewsML G2 descriptions enabled by logical consistency check;
- Enhanced search of news items enabled by logical inferences from the thesaurus and the knowledge formalized on the web;
- Unified semantic interfaces for searching and browsing seamlessly news content and background knowledge.

In the following, we describe the necessary steps for modeling such an ontology infrastructure. The various interconnected ontologies (NAR, NewsML-G2, EventsML-G2) are available at `http://newsml.cwi.nl/ontology/`.

### 4.1 Step 1: Modeling the NAR Ontology

The first step aims to capture formally the intended semantics of NAR and the family of IPTC G2 standards. Even though these models exist in UML diagrams, their "ontologisation" is not trivial. We discuss below the rationale of our modeling decisions.

[16] See the ReDeFer project: `http://rhizomik.net/redefer`

**Flattening the XML structure:** XML Schema provides the means to have very rich structure but is rather limited when expressing the meaning of this structure as the language is (only) concerned with providing typing and structuring information for isolated chunk of data. Consequently, the NAR model defines intermediate structures and *containers* whose only goal are to group a number of properties without particular semantics. These structures should not be represented in the ontology, as they will generate blank nodes in the RDF graph at the instance level, complexifying its visualization in any Semantic Web browser. While modeling the ontology, we therefore advocate to flatten the XML structure keeping only the properties that will be instantiated.

**Reification:** Statements about news items need often to be reified. For example, an editor registered as `team:md` can classify a news item as `diplomacy` at `2005-11-11T08:00:00Z`. Using the RDF reification and the N3 syntax, this yields the following statements:

```
{<> nar:subject cat:11002000} dc:creator   team:md ;
                              dc:modified  ‘‘2005-11-11T08:00:00Z’’ .
```

The RDF reification having no model theory semantics, we advocate the use of Networked Graphs where the relationships between graphs are described declaratively using SPARQL queries and an extension of the SPARQL semantics [16].

**Modeling unique identifiers** : News items metadata make use of numerous thesauri that implements a coding scheme for identifying the terms in order to be language agnostic. For example:

```
<pubStatus code="stat:usable"/>
<locCreated code="city:Paris"/>
<creator code="team:DOM"/>
<subject code="cat:04000000"/>
<subject code="isin:NL0000361939"/>
<subject code="pers:021147"/>
```

IPTC has therefore defined the notion of QCODES (by analogy to the XML QNAMES) with the following properties:

- Each coding scheme is associated with a URI. That URI must resolve to a resource (or resources) containing information about the scheme.
- The prefix represents the URI of the scheme within which the local part is allocated.
- There are almost no constraints on the values of the local part. For example, the local part (the code) is allowed to start with a digit.
- The two taken together must form a legal URI.
- This URI should provide access to a definition of the concept represented by that code within that scheme, i.e. it is dereferencable.

The tuple `prefix:localname` is however not identical to a CURIE[17] since the two parts (scheme and code) have each a meaning. For solving this issue, we advocate the "slash" rule, i.e. the concatenation of the scheme URI, a slash and the code, for the construction of a valid and dereferencable code URI.

### 4.2 Step 2: Linking with Media Ontologies

As we have discussed in the Section 2.2, other multimedia standards such as EXIF, Dublin Core, XMP, DIG35 or MPEG-7 are used in the media industry. These standards have generally been converted into OWL ontologies and can thus be integrated within our ontology infrastructure. Therefore, this step consists in adding OWL axioms stating the relationship between resources defined in different but strongly overlapping ontologies. For example, the NAR ontology contains the following axioms:

`nar:subject owl:equivalentProperty dc:subject`

`nar:Person owl:equivalentClass foaf:Person`

Semantic Web search engines such as Sindice[18], Watson[19] or Falcon[20] are useful tools for discovering concepts and properties defined in other ontologies that share the same semantics as the ones defined in our news infrastructure and could be linked to them.

### 4.3 Step 3: Converting IPTC News Codes into SKOS Thesaurus

The IPTC NewsCodes define 36 thesauri used as metadata values in the NAR architecture. Although the terms are sometimes organized in a taxonomy, the subsumption relationship is not explicit but instead encoded into the coding scheme identifying the terms. For example, "cancer" (`cat:07001004`) is narrower than "disease" (`cat:07001000`) which is narrower than "health" (`cat:07000000`) because they share a number of digits. We have converted these thesauri into SKOS, an application of RDF, making the subsumption relationships explicit (`skos:narrower`, `skos:broader`).

This RDF compatibility allows us to define some concepts in the NAR ontology in terms of `owl:Restriction`: the value of a property can be a `skos:Concept` or must come from a given `skos:ConceptScheme`. For example, the `nar:subject` object property is defined as having all its values from the IPTC Subject Codes `skos:ConceptScheme`.

Finally, we have exposed all these thesauri at `http://newsml.cwi.nl/NewsCodes/` following the Best Practice Recipes for Publishing RDF Vocabularies[21] and Cool URIs for the Semantic Web[22] notes. Each term is thus identified by a dereferencable URI. Consequently, sending an http request with the requested type

[17] `http://www.w3.org/TR/curie/`
[18] `http://sindice.com/`
[19] `http://watson.kmi.open.ac.uk/WatsonWUI/`
[20] `http://www.falcons.com.cn/`
[21] `http://www.w3.org/TR/swbp-vocab-pub/`
[22] `http://www.w3.org/TR/cooluris/`

`Accept:text/html` will deliver the original XML human readable version from IPTC of the thesauri, while the requested type `Accept:application/rdf+xml` will return the SKOS/RDF machine processable version of the thesaurus.

### 4.4 Step 4: Enriching the News Metadata

Once the NAR ontology has been modeled, linked to other media ontologies, and the thesauri converted into SKOS, the conversion of the metadata of individual news items into RDF according to this ontology infrastructure is straightforward. However, we advocate a further step aiming at enriching semantically the news metadata following the linked data principle[23]. In our case, we apply linguistic processing of textual news items and visual analysis of photo and video news items in order to extract more semantic metadata (Figure 1).

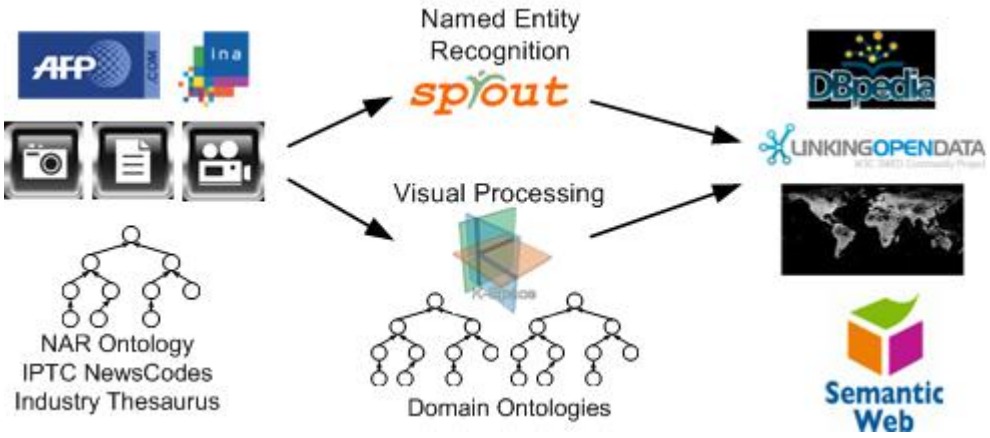


**Fig. 1.** Metadata enrichment of NewsItems

The linguistic processing consists in extracting named entities such as persons, organisations, locations, brands, etc. from the textual stories. Named Entity Recognizers such as GATE[24], SPROUT[25] or the most recent OpenCalais infrastructure[26] can be used. Once the named entities have been extracted, we map them to formalized knowledge on the web available in Geonames for the locations, or in DBPedia for the persons and organisations. Visual analysis provides additional metadata useful for organizing the results of a semantic news search engine. For example, an unsupervised clustering of photo news items can

[23] `http://linkeddata.org/`
[24] `http://gate.ac.uk/`
[25] `http://sprout.dfki.de/`
[26] `http://www.opencalais.com/`

be obtained using texture and color histograms, allowing to distinguish the photos of a football player such as Zinedine Zidane on the field, versus in a suit while he receives some award.

## 5 Semantic Search of Multimedia News

In order to demonstrate the appropriateness of our ontology infrastructure, we present an exploratory environment for searching and browsing news items. We use the Cliopatria[27] semantic search web-server [13, 21]. ClioPatria is a SWI-Prolog based platform for Semantic Web Applications that provides a scalable in-core RDF triple store and joins the SWI-Prolog RDF and HTTP infrastructure with a SeRQL/SPARQL query engine, interfacing to the The Yahoo! User Interface Library (YUI) and libraries that support semantic search. In contrast to client-only architectures such as Simile's Exhibit, ClioPatria has a client-server architecture. The core functionality is provided as HTTP APIs by the server. The results are served as presentation neutral data objects and can thus be combined with different presentation and interaction strategies. We have decided to use this open source software as it allows us to create customized presentations for searching and exploring news items, while being based on Semantic Web technologies and benefit from inference reasoning and SPARQL querying.

In the following, we present first the dataset used in our experiment (Section 5.1). We show then how we use the semantic metadata as dimensions for presenting the results of semantic search (Section 5.2) and for guiding a faceted browser like interface (Section 5.3) in the news domain.

### 5.1 Dataset

The dataset used in our experiment consists of the ontology infrastructure detailed previously, 60000 news stories in English, 40000 news stories in French, 2557 photos and 8 hours of broadcasted video (Table 1).

Following the four steps detailed in the Section 4, we have processed the news items in order to enrich the metadata. We have used SPROUT together with a specific football gazetteer in order to extract named entities from the caption of the 2557 photos contained in our dataset. The use of a domain specific ontology allows us to extract more semantic information such as the role of a football player (goalkeeper, middlefielder), the name of a team, etc. The Figure 2 (resp. 3) shows the algorithm for linking the entities of type `Person` (resp. `Location`) with DBPedia (resp. Geonames).

This processing step provided 217 DBPedia persons and 426 Geonames locations. The assessment of the results shows that the Geonames web service tend to return primarily a US city when a single string is passed as an argument. Fortunately, news items contain always information about the city and the country yielding accurate recognition of the location mentioned in the story. The few

[27] `http://e-culture.multimedian.nl/software/ClioPatria.shtml`

errors we noticed come from an incorrect typing of the named entity from the SPROUT system, e.g. `Australia` as been typed as a `Person`. More sophisticated disambiguation heuristics such as IdentityRank [8] can be further employed to minimize these errors.

| Description | Number of RDF triples |
|---|---|
| General ontologies: NAR, NewsML-G2, DC, VRA, FOAF | 7,336 |
| Domain specific ontologies: Football ontology | 104,358 |
| Thesauri: IPTC NewsCodes, INA Thesaurus | 34,903 |
| External resources: Geonames, DBPedia | 53,468 |
| AFP News Feed in English from June and July 2006 | 804,446 |
| AFP Photos from the 2006 World Cup | 61,311 |
| INA Broadcast Video from June and July 2006 | 1,932 |
| **Total** | 1,067.754 |

**Table 1.** Number of RDF triples loaded in our semantic search web-server

### 5.2 Semantic Search of News Items

The Figure 4 shows the result for the query "Lyon" in our semantic search system. The news items are grouped according to the path in the RDF graph that leads to the property for which the value has matched the query. In our case, the system returns the news items where "Lyon" occurs in the `title`, the `headline`, the `slugline`, etc. Each group can be collapsed or expanded. Furthermore, the information about the type of news item allows us to customize the visual rendering of each group: text news items have a snippet view displaying the first three lines of the stories, while photo news items are displayed in a thumbnail carousel.

---

For each named entity of type `Person` recognized, do:

1. Construct a SPARQL query for DBPedia using the `rdfs:label` property and all supported languages and return the first resource
2. Construct a SPARQL query for the Football ontology using the `dolce:firstName` and `dolce:lastName` properties and return the first resource
3. If a resource is found both in DBPedia and in the Football ontology, then add a `owl:sameAs` statement between these two resources
4. If no resource is found in DBPedia and in the Football ontology, then create a new instance of `Person` in the knowledge base

---

**Fig. 2.** Pseudo algorithm for linking the extracted named entities with DBPedia

Interestingly, the last group of result contain a single news photo item depicting three football players. The metadata of this photo does not contain the

string “Lyon”. Instead, the caption of this photo mentions Juninho Pernambucano, recognized by SPROUT as a football player, later on linked with the DBPedia resource identifying this person. DBPedia contains information about this person such as his birthdate and all the past teams where he played. Among them, “Lyon” is his current club and this is why this image has been retrieved, even though at the end of the list because of the length in the RDF graph necessary to reach this term.

---

For each named entity of type `Location` recognized, do:

1. Get the location and if available the broader location
2. Construct a textual query for Geonames with either the exact location name or with both terms and return the first resource, for example:
   `http://ws.geonames.org/search?maxRows=1&type=rdf&name_equals=Germany`
   or `http://ws.geonames.org/search?maxRows=1&type=rdf&q=Berlin,Germany`

---

**Fig. 3.** Pseudo algorithm for linking the extracted named entities with Geonames


**Fig. 4.** Search for “Lyon” in the semantic search engine

Similarly, the query for “Saksamaa” returns a single group of 679 photos while none of them contain this term in the metadata. Again, the explanation is that all these photos have been captured in Germany (during the World Cup), a named entity recognized by SPROUT as a location, later on linked with Geonames.

The Geonames resource contains information about all the alternative names of Germany in all languages, Saksamaa being the Ethiopian name of Germany.

### 5.3 Semantic Browsing of Multimedia News

Additionally to the semantic search interface, we provide a faceted browser like interface for better exploring the news dataset. Facets correspond to properties of interest in the metadata, and can be selected by the end-user. We have defined a soccer view gathering the properties `subject`, `slugline`, `locCreated`, `location` and `person`. Using the information provided by Geonames, we are able to propose more views for presenting the information. The Figure 5 shows such a view, where the football player `Zinedine Zidane` has been selected as a filter: the red flags correspond then to the countries mentioned in the news stories, while the blue flags correspond to the cities where the news stories have been produced.

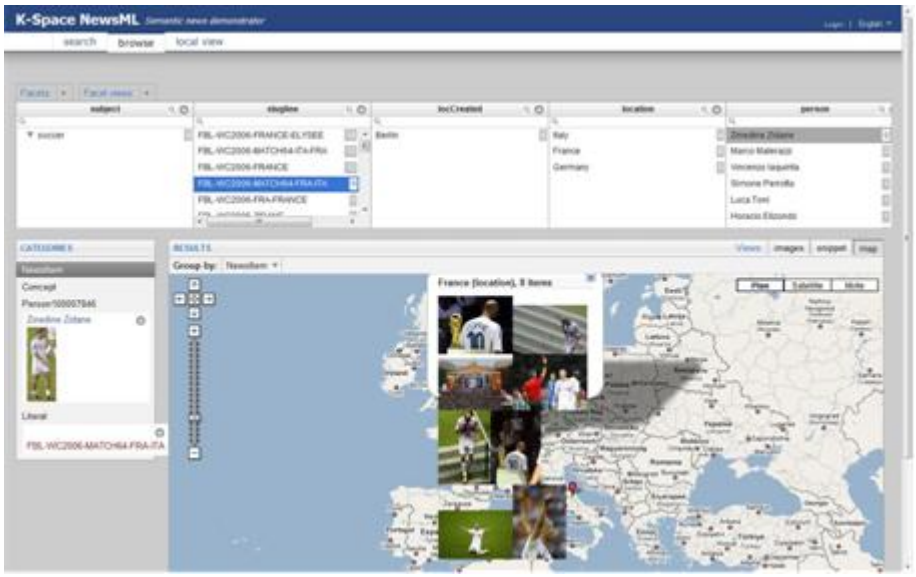


**Fig. 5.** Browsing the photos captured in France during the World Cup finale

Interestingly, this view allows the end-user to immediately distinguish between the two sluglines `FBL-WC2006-MATCH64-ITA-FRA` and `FBL-WC2006-MATCH64-FRA-ITA` that look really similar but actually correspond to the news stories produced in Italy (Italian point of view) and in France (French point of view). Such a subtlety is hard to see in the metadata while the map view gives an immediate insight of the data.

Finally, the Figure 6 shows the local view of a video resource. The metadata corresponds to a particular sequence in a TV news broadcast program. Arbitrary

sequences of a video can be played in the semantic browser using the `tcin` and `tcout` buttons. An auto-play has been considered but the non-ability of the current web infrastructure to address temporal fragments of a video file prevent such a functionality [19].


**Fig. 6.** Video local view: arbitrary temporal segment can be played

## 6 Conclusion and Future Work

In this paper, we have described a method composed of four steps for building an ontology-based news infrastructure. These guidelines are complementary to the development of ontology design patterns and best practices for publishing semantic web vocabularies that is central in the web of data. We have discussed the design decisions regarding the modeling of the NAR ontology from existing XML Schemas. At the ontology level, we advocate to flatten the XML structure, to identify properly the resources and to reuse as much as possible existing ontologies. At the instance level, we recommend to enrich and link the metadata with existing SKOS thesauri and formalized knowledge existing on the web such as DBPedia and Geonames. The NAR ontology is currently reviewed by the IPTC and could be endorsed by the standardisation body. We presented a semantic search system and various exploratory interfaces for searching and browsing news items. These interfaces use the richness of the semantic metadata for grouping, ranking and presenting the results of a given query. The system is publicly available at `http://newsml.cwi.nl/explore/search`.

Time is an essential dimension in the news domain and our system provides also a timeline view. Nevertheless, reasoning on time information is a complex task. From the representation point of view, we plan to include the Time Ontology[28] and the temporal relations module from the DOLCE upper ontology in order to propose histogram views aggregating the stories per topic and per day, week or month. Our current system works on static, pre-processed and staged data. A natural evolution is to create a dynamic environment where incoming news feed is processed in live and immediately available to the end-user. Finally, an evaluation of the system by AFP journalists is planned.

## Acknowledgments

The dataset used has been kindly provided by AFP for the news stories and the photos, and by INA for the videos. The author would like to particularly thank Laurent Le Meur from AFP for fruitful discussions on the design of the NAR ontology. The author would also like to thank the following colleagues at CWI, Amsterdam (Lynda Hardman, Michiel Hildebrand, Michiel Kauw-A-Tjoe, Zeljko Obrenovic and Jacco van Ossenbruggen) and at IBBT Multimedia Lab, Ghent University (Erik Mannens, Gaëtan Martens and Chris Poppe) for their feedback on the prototype and earlier versions of this paper. The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space.

## References

1. R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In *6th International Semantic Web Conference (ISWC'07)*, pages 30–43, Busan, Korea, 2007.
2. M. van Assem, V. Malaisé, A. Miles, and G. Schreiber. A Method to Convert Thesauri to SKOS. In *3rd European Semantic Web Conference (ESWC'06)*, pages 95–109, Budva, Montenegro, 2006.
3. M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A Method for Converting Thesauri to RDF/OWL. In *3rd International Semantic Web Conference (ISWC'04)*, pages 17–31, Hiroshima, Japan, 2004.
4. B. Bachimont, A. Isaac, and R. Troncy. Semantic Commitment for Designing Ontologies: A Proposal. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, pages 114–121, Sigüenza, Spain, 2002.
5. P. Castells, F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras, and J. Lorés. Neptuno: Semantic Web Technologies for a Digital Newspaper Archive. In *1st European Semantic Web Symposium (ESWS'04)*, pages 445–458, Heraklion, Crete, 2004.
6. M. Fernández, A. Gómez-Pérez, and N. Juristo. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In *AAAI97 Spring Symposium Series on Ontological Engineering*, pages 33–40, Stanford, California, USA, 1997.

[28] `http://www.w3.org/TR/owl-time/`

7. N. Fernández, J. M. Blázquez, J. Arias, L. Sánchez, M. Sintek, A. Bernardi, M. Fuentes, A. Marrara, and Z. Ben-Asher. NEWS: Bringing Semantic Web Technologies into News Agencies. In *5th International Semantic Web Conference (ISWC'06)*, pages 778–791, Athens, Georgia, USA, 2006.
8. N. Fernández, J. M. Blázquez, L. Sánchez, and A. Bernardi. IdentityRank: Named Entity Disambiguation in the Context of the NEWS Project. In *4th European Semantic Web Conference (ESWC'07)*, pages 640–657, Innsbruck, Austria, 2007.
9. N. Fernández, L. Sánchez, J. M. Blázquez, and J. Villamor. The NEWS Ontology for Professional Journalism Applications. In *A Handbook of Principles, Concepts and Applications in Information Systems*, volume 14 of *Integrated Series in Information Systems*. Springer, 2007.
10. R. Garcia, F. Perdrix, R. Gil, and M. Oliva. The semantic web as a newspaper media convergence facilitator. *Journal of Web Semantics*, 6(2):151–161, 2008.
11. A. Gómez-Pérez, M. Fernandez-Lopez, and Oscar Corcho. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Series: Advanced Information and Knowledge Processing. Springer, first edition, 2004.
12. M. Hausenblas, S. Boll, T. Bürger, O. Celma, C. Halaschek-Wiener, E. Mannens, and R. Troncy. Multimedia Vocabularies on the Semantic Web. W3C Multimedia Semantics Incubator Group Report, 2007. `http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies/`.
13. M. Hildebrand, J. van Ossenbruggen, and L. Hardman. /facet: A Browser for Heterogeneous Semantic Web Repositories. In *5th International Semantic Web Conference (ISWC'06)*, pages 272–285, Athens, Georgia, USA, 2006.
14. H. Knublauch, D. Oberle, P. Tetlow, and E. Wallace. A Semantic Web Primer for Object-Oriented Software Developers. W3C Note, 2006. `http://www.w3.org/TR/sw-oosd-primer/`.
15. MPEG-7. Multimedia Content Description Interface. ISO/IEC 15938, 2001.
16. S. Schenk and S. Staab. Networked Graphs: A Declarative Mechanism for SPARQL Rules, SPARQL Views and RDF Data Integration on the Web. In *17th International World Wide Web Conference (WWW'08)*, Beijing, China, 2008.
17. A. Tordai, B. Omelayenko, and G. Schreiber. Semantic Excavation of the City of Books. In *Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM'07)*, pages 39–46, 2007.
18. R. Troncy, Ó. Celma, S. Little, R. García, and C. Tsinaraki. MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In *1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies (MAReSO)*, Genova, Italy, 2007.
19. R. Troncy, L. Hardman, J. van Ossenbruggen, and M. Hausenblas. Identifying Spatial and Temporal Media Fragments on the Web. In *W3C Video on the Web Workshop*, 2007.
20. M. Uschold and M. Grüninger. Ontologies: Principles, Methods and Applications. *Knowledge Engineering Review*, 2:93–155, 1996.
21. J. Wielemaker, M. Hildebrand, J. van Ossenbruggen, and G. Schreiber. Thesaurus-based search in large heterogeneous collections. In *7th International Semantic Web Conference (ISWC'08)*, Karlsruhe, Germany, 2008.
22. B. Wielinga, J. Wielemaker, G. Schreiber, and M. van Assem. Methods for Porting Resources to the Semantic Web. In *1st European Semantic Web Symposium (ESWS'04)*, pages 299–311, Heraklion, Greece, 2004.