The Semantic Web as a Newspaper Media Convergence Facilitator

Roberto García[1*], Ferran Perdrix[1,2], Rosa Gil[1], Marta Oliva[1]

[1]Departament d'Informàtica i Enginyeria Industrial,
Universitat de Lleida,
Jaume II 69, E-25001 Lleida, Spain
Fax +34 973 702 702
{rgarcia, fperdrix, rgil, oliva}@diei.udl.cat

[2]Diari Segre Media Group,
Del Riu, 6, E-25007 Lleida, Spain
fperdrix@diarisegre.com

Abstract

Newspapers are evolving and this causes great changes in how newspapers reach their consumers, but also in how newspapers work internally. Advanced computerised support is needed in order to cope with the new needs, which require that machines are aware of a greater part of the underlying semantics. Ontologies and Semantic Web technologies are clear candidates for web-wide semantics. However, newspapers have made great investments in their current news management systems and their wish is to undertake a smooth transition. Our proposal is to build an ontological framework based on existing journalism and multimedia standards. These standards are based on XML technologies. Therefore, we have developed a generic XML Schema to OWL mapping, complemented with an XML to RDF one. Together, they allow reusing existing metadata that, once in a semantic space, facilitates data integration, news management and retrieval. The resulting ontological framework is being applied in the Diari Segre Media Group[1], which produces press, radio and television content.

Keywords

News, newspaper, ontology, multimedia, integration.

1 Introduction

Web news publishing is evolving fast, as the majority of Internet services, and nowadays this service is trying to adapt information to a way that best fits user's interests in order to increase its usage. With that, newspapers are looking forward to profit more from their news sites. In parallel, many of the newspaper companies are changing into news media houses. They own radio stations and video production companies that produce content unsupported by traditional newspapers, but that is delivered by Internet newspapers or new mobile services. Initially, Internet news were a mere reproduction of those in the printed edition. Nowadays, Internet news are evolving fast, they are constantly updated and provide new services for those users interested on reaching this information as soon as possible and enjoying new ways of interaction with them [1,2,3].

Consequently, news industry communication model is changing from the traditional one shown on the left of Fig. 1 to the one shown in the right. In the former, each channel is considered separately (press, TV, radio, Internet, mobile phones…) and implies his way creating his own message, transmitting over this channel and using his own interface in order to show the message to the receivers. On the other hand, the latter is based on an information convergence flux. In this model, transmitters make information in collaboration with other transmitters and produce messages that include as media as it is necessary (video, text, audio, images…). Finally, receivers

[1] http://www.diarisegre.com

choose the channel that best fits their needs in order to get access to messages. Moreover, the involved roles, transmitter and receiver, are not fixed as before. Newspapers promote that their readers become also transmitters. In any case, the new content flux generalises both situations as roles do not change, just who plays them.

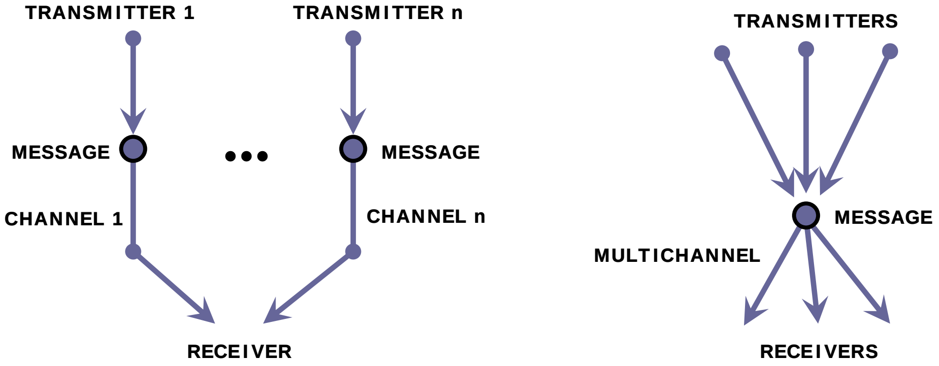


Fig. 1. Traditional news information flux (left) and new convergent flux (right)

Media houses must adapt to the requirements imposed by this new model. First of all, there are changes in how they reach consumers. News are build up from a combination of different content types (video, audio, the traditional text plus images, etc.) and are delivered to users through different channels and adapted to many kinds of devices (PC, PDA, smart phones, etc.). Therefore, formats must be selected and adapted according to the device and connection the user is using. These operations include transcoding of formats, resizing of images or recoding for higher levels of compression. Moreover, multi-channel distribution must take into account that for each channel we must define its own content, aesthetic and interaction model. These characteristics define what an interactive channel is [4].

However, changes are not just restricted to the relation with consumers. Digital media eliminates many time and space restrictions and changes editorial team routines. Moreover, all different media converge into a unified news object that is produced by interdisciplinary teams. Consequently, more efficient and effective means for news management are needed in order to facilitate coordination and production of these multimedia assets.

The news industry is currently using content management solutions for these means, but the additional requirements of a convergent editorial office stress the need for advanced knowledge management and information retrieval. Currently, knowledge management and information retrieval is carried out mainly by the documentation department, who is in charge of the news archival process. They classify news using a hierarchical thesaurus.

Journalists search this information when they need to inform themselves on subjects, stories or events. This search can be performed in extreme situations, e.g., lack of time, or lack of knowledge in relation to the archive system. This is reflected in the way journalists formulate their queries. The gap between archivists' and journalists' mental models implies that more flexible content categorization and search systems are needed. For instance, in the Segre media house, archivists are used to the proprietary content taxonomy that has been used for years while journalist prefer the standard taxonomy proposed by IPTC, IPTC NewsCodes[2]. This trend is even bigger when we consider cross-media content production and coordination in order to get multimedia news.

Therefore, in order to take advantage of the possibilities offered by the digital medium to exploit a newspaper archive, the aspects that can be improved include:

- keyword search falling short in expressive power;
- weak interrelation between archive items: users may need to combine several indirect queries manually before they can get answers to complex queries;
- lack of a commonly adopted standard representation for sharing archive news across newspapers;
- lack of internal consensus for content description terminology among reporters and archivists;
- lack of involvement of journalist in the archiving process.

In this paper, we explore Semantic Web technologies [5] as a way to overcome many of the challenges of digital and convergent media houses. The size and complexity of the stored

[2] NewsCodes, http://www.iptc.org/NewsCodes

information, and the time limitations for cataloguing, describing and ordering the incoming information, make newspaper archives a relatively disorganised and difficult to manage corpus. In this sense, they share many of the characteristics and problems of the WWW, and therefore the solutions proposed in the Semantic Web vision are pertinent here.

In order to implement more advanced newspaper content management applications, they should be more informed about the content they are managing. They are not just files with some weak interrelations. There is a lot of knowledge embedded in these pieces of content and in their interrelationships. In order to make computers aware of it, the implicit semantics must be formalised using ontologies, the knowledge engineering artefact for computerised knowledge representation and management. Semantic Web technologies facilitate the building block for Web ontologies, which add the facilities for web-wide ontology sharing and integration. The latter is a key feature for convergent and globalised media houses.

The rest of this paper is organised as follows. Section 2 presents the state of the art of journalism and multimedia metadata, and then analyses the current situation. Section 3 presents the methodology that has been employed in order to produce an ontological framework based on existing standards. Section 4 shows the ontological infrastructure that has been generated, the architecture of the semantics processing infrastructure and some examples of semantic integration mappings. In Section 5, one of the applications that have been developed on top of the Semantic Web infrastructure is introduced. Finally, there are the conclusions and a sketch of the future plans in Section 6.

2 State of the Art

In order to build an ontological infrastructure for the Semantic Newspaper, it is important to analyse the state of the art of the metadata initiatives in the journalism domain. Additionally, digital newspapers have stressed the requirements of multimedia management. Digital news are managed as multimedia packages that integrate text, images, video, audio, etc. Therefore, it is also important to analyse the current situation in the more general multimedia metadata domain. Both aspects are studied next, first journalism metadata and then multimedia metadata. This section ends with an analysis of the current situation, which identifies the keys points that have guided our approach to build the ontological framework.

2.1 Journalism Metadata

One of the main standardization frameworks in the journalism domain is the International Press Telecommunications Council[3] (IPTC), an international consortium of news agencies, editors and newspapers distributors. IPTC has developed standards like the Information Interchange Model[4] (IIM), NewsCodes (formerly the Subject Reference System), the News Industry Text Format[5] (NITF) or NewsML[6].

Currently, almost all of them have evolved towards XML-based standards to represent and manage news along their whole lifecycle, including their creation, exchange and consumption.

For instance, NewsML is used to represent news as multimedia packages and NITF deals with document structure, i.e. paragraphs, headlines, etc. On the other hand, the Subject Reference System (SRS), now part of IPTC NewsCodes, is a subject classification hierarchy with three levels and seventeen categories in its first level.

This kind of metadata constitutes one of the kinds managed by Segre Media Group systems. It is used for items from different news agencies and currently constitutes one of the main integration efforts.

[3] IPTC, http://www.iptc.org

[4] IIM, http://www.iptc.org/IIM

[5] NITF, http://www.nitf.org

[6] NewsML, http://www.newsml.org

2.2 Multimedia News Items Metadata

All the previous initiatives are centred on the journalism specific aspects of a semantic newspaper. However, as it has been pointed out, newspapers are evolving towards the digital multimedia domain. Therefore, they stress more and more their multimedia management requirements.

In the multimedia metadata domain, as it is extensively shown in the literature [6,7,8,9], the MPEG-7 [10] standard constitutes the greatest effort for multimedia description. It is divided into four main components: the Description Definition Language (DDL, the basic building blocks for the MPEG-7 metadata language), Audio (the descriptive elements for audio), Visual (those for video) and the Multimedia Description Schemes (MDS, the descriptors for capturing the semantic aspects of multimedia contents, e.g. places, actors, objects, events, etc.).

In addition to MPEG-7, which concentrates on content description, MPEG-21 defines an open framework for multimedia delivery and consumption. This standard must be also considered because it focuses on the content management issues for full delivery and consumption chain, from content creators' applications to end-users players. The different MPEG-21 parts deal with diverse aspects like Digital Rights Management or Digital Items, the definition of a fundamental content unit for distribution and transaction very useful for convergent media management. This common framework might help integrating different multimedia metadata schemes, especially proprietary ones like the one provided by the Milenium Quay[7] cross-media archive system currently being used at Segre Media Group.

2.3 Situation Analysis

The main standards that have been presented, both in the journalism and multimedia domains, are based on XML and specified by XML Schemas. The more significant case is the MPEG-7 one. It is based on a set of XML Schemas that define 1182 elements, 417 attributes and 377 complex types. NewsML and NITF are also very big standards, they define more than 100 elements, and the NewsCodes hierarchy of subjects defines more than one thousand different subjects.

The size of these standards makes it very difficult to manage them. Moreover, the use of XML technologies implies that a great part of the semantics remains implicit. Therefore, each time an application is developed; semantics must be extracted from the standard and re-implemented.

MPEG-7 constitutes a valuable starting point for more specific developments, i.e. it can be seen as an "upper-ontology" for multimedia. However, the lack of explicit semantics makes MPEG-7 very difficult to extend in an independent way, i.e. third party extensions.

The same applies for MPEG-21 or the journalism standards. Moreover, standards from both worlds share many concepts so it would be possible, and easier, to integrate them once their implicit semantics are available from a computer processing point of view.

Our approach to the inconveniencies observed in this state of the art is presented in the next section.

3 Approach

We have undertaken the application of the Semantic Web proposals to the newspapers world by following a smooth transition strategy [11]. This strategy advises about keeping compatibility (at least initially) with the current technology: browsers, protocols, web and application servers, databases, and architectures.

The objective is then to design a platform that is an extension of previously working systems in mass media companies. The manual creation of semantic instances for news items, at a regular daily pace, is indeed a feasible goal in the future. The introduction of new semantic documentation tools requires, however, a careful work of analysis, design, testing and balancing of the additional burden that such tools may impose on archivists, journalists or end-users.

In order to put into practice the smooth transition strategy, the first step has been to reuse existing standards in the journalism and multimedia fields, which have been for long very active in standardization.

However, as it has been highlighted in the state of the art, all the more recent standards are based on XML but lack formal semantics that facilitate applying a Semantic Web approach.

[7] Milenium Quay, http://www.mileniumcrossmedia.com

Therefore, in order to facilitate the transition from current standards and applications to the semantic world, we have applied the XML Semantics Reuse methodology, which is detailed in section 3.2.

3.1 Existing Approaches

There are other existing initiatives that try to move journalism and multimedia metadata to the Semantic Web world. In the journalism field, the Neptuno [12] and NEWS [13] projects can be highlighted. Both projects have developed ontologies based on existing standards (IPTC SRS, NITF or NewsML) but from an ad-hoc point of view. Therefore, in order to smooth the transition from the previous legacy systems, complex mappings should be developed and maintained.

The same can be said for the existing attempts to produce semantic multimedia metadata. Chronologically, the first attempts to make MPEG-7 metadata semantics explicit where carried out, during the MPEG-7 standardisation process, by Jane Hunter [14]. The proposal used RDF to formalise a small part of MPEG-7, and later incorporated some DAML+OIL construct to further detail their semantics [15]. However, at that moment, there were not mature technologies for Web-wide metadata semantics formalisation. Moreover, XML had already a great momentum, so it was the logical choice.

From this point, once Semantic Web has matured, there have been more attempts to relate MPEG-7 with Web ontologies. However, none of them has retaken the initial effort to completely move MPEG-7 to the Semantic Web. This initiatives range from separated modules for existing MPEG-7 tools that offer reasoning capabilities for concrete aspects of multimedia management [7], to a partial OWL modelling of the MPEG-7 Multimedia Description Schemes intended to facilitate MPEG-7 extensions [9]. Moreover, they are not systematic; they are applied on an ad-hoc basis, which makes them very costly to apply to the whole MPEG-7 standard.

All these initiatives have produced very interesting results, both in the journalism and multimedia fields, and they are complementary to our objective, i.e. to smoothly move towards a semantic newspaper based on the main journalism and multimedia standards.

The method we have used to perform this is detailed in the next section. It is a generic XML Schema to OWL mapper combined with an XML to RDF translator. It has already shown its usefulness with other quite big XML Schemas in the Digital Rights Management domain, such as MPEG-21 REL [16] and ODRL [17], and in the music metadata field [18].

3.2 XML Semantics Reuse Methodology

The main caveat of semantic multimedia metadata is that it is sparse and expensive to produce. The previously introduced initiatives are appropriate when applied to limited scopes. However, if we want to increase the availability of semantic multimedia metadata and, in general, of semantic metadata, we need methods that are more productive. The more direct solution is to take profit from the great amount of metadata that has been already produced by the XML community.

There are many attempts to move metadata from the XML domain to the Semantic Web. Some of them just model the XML tree using the RDF primitives [19]. Others concentrate on modelling the knowledge implicit in XML languages definitions, i.e. DTDs or the XML Schemas, using web ontology languages [20,21]. Finally, there are attempts to encode XML semantics integrating RDF into XML documents [22,23].

However, none of them facilitates an extensive transfer of XML metadata to the Semantic Web in a general and transparent way. Their main problem is that the XML Schema implicit semantics are not made explicit when XML metadata instantiating this schemas is mapped. Therefore, they do not take profit from the XML semantics and produce RDF metadata almost as semantics-blind as the original XML. Or, on the other hand, they capture these semantics but they use additional ad-hoc semantic constructs that produce less transparent metadata.

Therefore, we have chosen the XML Semantics Reuse methodology [24] and the XML Schema to OWL and XML to RDF tools implemented in the ReDeFer project[8]. This methodology combines an XML Schema to web ontology mapping, called XSD2OWL, with a transparent mapping from XML to RDF, XML2RDF. The ontologies generated by XSD2OWL are used

[8] ReDeFer project, http://rhizomik.net/redefer

during the XML to RDF step in order to generate semantic metadata that makes XML Schema semantics explicit. Both steps are detailed next.

XSD2OWL Mapping

The XML Schema to OWL mapping is responsible for capturing the schema implicit semantics. This semantics are determined by the combination of XML Schema constructs. The mapping is based on translating these constructs to the OWL ones that best capture their semantics. These translations are detailed in Table 1.

The XSD2OWL mapping is quite transparent and captures a great part XML Schema semantics. The same names used for XML constructs are used for OWL ones, although in the new namespace defined for the ontology. XSD and OWL constructs names are identical; this usually produces uppercase-named OWL properties because the corresponding element name is uppercase, although this is not the usual convention in OWL.

Therefore, XSD2OWL produces OWL ontologies that make explicit the semantics of the corresponding XML Schemas. The only caveats are the implicit order conveyed by *xsd:sequence* and the exclusivity of *xsd:choice*.

For the first problem, *owl:intersectionOf* does not retain its operands order, there is no clear solution that retains the great level of transparency that has been achieved. The use of RDF Lists might impose order but introduces ad-hoc constructs not present in the original metadata. Moreover, as it has been demonstrated in the Semantic Web community, the element ordering does not contribute much from a semantic and knowledge representation point of view [25]. For the second problem, *owl:unionOf* is an inclusive union, the solution is to use the disjointness OWL construct, *owl:disjointWith,* between all union operands in order to make it exclusive.

Table 1. XSD2OWL translations for the XML Schema constructs and shared semantics with OWL constructs

XML Schema	OWL	Shared informal semantics
element | attribute	rdf:Property owl:DatatypeProperty owl:ObjectProperty	Named relation between nodes or nodes and values
element@substitutionGroup	rdfs:subPropertyOf	Relation can appear in place of a more general one
element@type	rdfs:range	The relation range kind
complexType|group |attributeGroup	owl:Class	Relations and contextual restrictions package
complexType//element	owl:Restriction	Contextualised restriction of a relation
extension@base | restriction@base	rdfs:subClassOf	Package concretises the base package
@maxOccurs @minOccurs	owl:maxCardinality owl:minCardinality	Restrict the number of occurrences of a relation
sequence choice	owl:intersectionOf owl:unionOf	Combination of relations in a context

To conclude, one important aspect is that the resulting OWL ontology may be OWL-Full. This is due to the fact that, in some cases, the XSD2OWL translator must employ *rdf:Property* for those *xsd:elements* that have both data type and object type ranges. However, this can be fixed in the case of MPEG-7 because there are only 23 elements that have both complex type and simple type content, which causes them to be mapped to *rdf:Properties*. Two different properties are generated, a *DatatypeProperty* and an *ObjectProperty,* for each of them. Consequently, the ontology can be made OWL-DL.

XML2RDF Mapping

Once all the metadata XML Schemas are available as mapped OWL ontologies, it is time to map the XML metadata that instantiates them. The intention is to produce RDF metadata as transparently as possible. Therefore, a structure-mapping approach has been selected [19]. It is also possible to take a model-mapping approach [26].

XML model-mapping is based on representing the XML information set using semantic tools. This approach is better when XML metadata is semantically exploited for concrete purposes. However, when the objective is semantic metadata that can be easily integrated, it is better to take a more transparent approach.

Transparency is achieved in structure-mapping models because they only try to represent the XML metadata structure, i.e. a tree, using RDF. The RDF model is based on the graph so it is easy to model a tree using it. Moreover, we do not need to worry about the semantics loose produced by structure-mapping. We have formalised the underlying semantics into the corresponding ontologies and we will attach them to RDF metadata using the instantiation relation *rdf:type*.

The structure-mapping is based on translating XML metadata instances to RDF ones that instantiate the corresponding constructs in OWL. The more basic translation is between relation instances, from *xsd:elements* and *xsd:attributes* to *rdf:Properties*. Concretely, *owl:ObjectProperties* for node to node relations and *owl:DatatypeProperties* for node to values relations.

However, in some cases, it would be necessary to use *rdf:Properties* for *xsd:elements* that have both data type and object type values. Values are kept during the translation as simple types and RDF blank nodes are introduced in the RDF model in order to serve as source and destination for properties. They will remain blank for the moment until they are enriched with semantic information.

The resulting RDF graph model contains all that we can obtain from the XML tree. It is already semantically enriched thanks to the *rdf:type* relation that connects each RDF property to the *owl:ObjectProperty* or *owl:DatatypeProperty* it instantiates. It can be enriched further if the blank nodes are related to the *owl:Class* that defines the package of properties and associated restrictions they contain, i.e. the corresponding *xsd:complexType*. This semantic decoration of the graph is formalised using *rdf:type* relations from blank nodes to the corresponding OWL classes.

At this point we have obtained a semantically enabled representation of the input metadata. The instantiation relations can now be used to apply OWL semantics to metadata. Therefore, the semantics derived from further enrichments of the ontologies, e.g. integration links between different ontologies or semantic rules, are automatically propagated to instance metadata thanks to inference.

However, before continuing to the next section, it is important to point out that these mappings have been validated in different ways. First, we have used OWL validators in order to check the resulting ontologies, not just the MPEG-7 Ontology but also many others [16,17]. Second, our MPEG-7 ontology has been compared to Jane Hunter's and Chrisa Tsinaraki's ones [15,9]. This comparison has shown that our mapping captures the same semantics as those captured by hand by Jane Hunter using RDF Schema and DAML+OIL. Moreover, the same OWL constructs as those used by Chrisa Tsinaraki are generated, plus many more that are difficult to capture without the help of computerised means. For instance, AudioType is modelled using a cardinality restriction on Audio but it lacks an all values from AudioSegmentType restriction on the same property. Finally, the two mappings have been tested in conjunction. Testing XML instances have been mapped to RDF, guided by the corresponding OWL ontologies from the used XML Schemas, and then back to XML. Then, the original and derived XML instances have been compared using their canonical version in order to correct mapping problems.

4 Ontological Infrastructure

As a result of applying the XML Semantics Reuse methodology, we have obtained a set of ontologies that reuse the semantics of the underlying standards, as they are formalised through the corresponding XML Schemas. All the ontologies related to journalism standards, i.e. NewsCodes NITF and NewsML, are available from the Semantic Newspaper site[9]. This site also contains some of the ontologies for MPEG-21 useful for news modelling as convergent multimedia units. The MPEG-7 Ontology is available from the MPEG-7 Ontology site[10].

The ontologies that are going to be used as the basis for the info-structure of the semantic newspaper are:

- **NewsCodes Subjects Ontology**: an OWL ontology for the subjects' part of the IPTC NewsCodes. It is a simple taxonomy of subjects. An existing RDF Schema from the Neptuno

[9] Semantic Newspaper, http://rhizomik.net/semanticnewspaper

[10] MPEG-7 Ontology, http://rhizomik.net/mpeg7ontos

project [12] has been reused. It has been trivially modified in order to make it an OWL ontology in order to facilitate the integration of the subjects' taxonomy in the global ontological framework. However, in order to avoid some undesired inferences due to the fact that the subjects are taxonomy terms and not classes, the future work is to model this taxonomy using SKOS [27].

- **NITF 3.3 Ontology**: an OWL ontology that captures the semantics of the XML Schema specification of the NITF standard. It contains some classes and many properties dealing with document structure, i.e. paragraphs, subheadlines, etc., but also some metadata properties about copyright, authorship, issue dates, etc.
- **NewsML 1.2 Ontology**: the OWL ontology resulting from mapping the NewsML 1.2 XML Schema. Basically, it includes a set of properties useful to define the news structure as a multimedia package, i.e. news envelope, components, items, etc.
- **MPEG-7 Ontology**: The XSD2OWL mapping has been applied to the MPEG-7 XML Schemas producing an ontology that has 2372 classes and 975 properties, which are targeted towards describing multimedia at all detail levels, from content based descriptors to semantic ones.
- **MPEG-21 Digital Item Ontologies**: a Digital Item (DI) is defined as the fundamental unit for distribution and transaction in MPEG-21. In order to model DIs, the MPEG-7 schemas for the Digital Item Identifier (DII), the Digital Item Declaration Model (DIDM) and the Digital Item Declaration Language (DIDL) have been mapped to OWL. Other parts of the MPEG-21 standard, specially the Rights Expression Language (REL) [16], have been also mapped to OWL using XSD2OWL and can be used if required.

4.1 System architecture

Based on the previous XML world to Semantic Web domain mappings, we have built up a system architecture that facilitates journalism and multimedia metadata integration and retrieval. The architecture is sketched in Fig. 2. The MPEG-7 OWL ontology, generated by XSD2OWL, constitutes the basic ontological framework for semantic multimedia metadata integration and appears at the centre of the architecture. In parallel, there are the journalism ontologies. The multimedia related concepts from the journalism ontologies are connected to the MPEG-7 ontology, which acts as an upper ontology for multimedia. Other ontologies and XML Schemas can also be easily incorporated using the XSD2OWL module.

Semantic metadata can be directly fed into the system together with XML metadata, which is made semantic using the XML2RDF module. For instance, XML MPEG-7 metadata has a great importance because it is commonly used for low-level visual and audio content descriptors automatically extracted from its underlying signals. This kind of metadata can be used as the basis for audio and video description and retrieval.

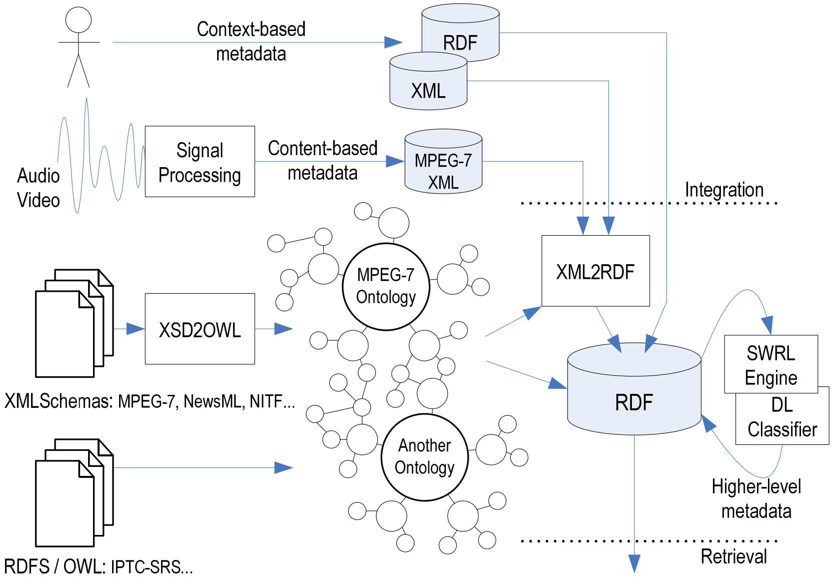


Fig. 2. Metadata integration and retrieval architecture

In addition to content-based metadata, there is context-based metadata. This kind of metadata is higher level and it is usually, in this context, related to journalism metadata. It is generated by the system users (journalist, photographers, cameramen, etc.). For instance, there are issue dates, news subjects, titles, authors, etc.

This kind of metadata can come directly from semantic sources but, usually, it is going to come from legacy XML sources based on the standards' XML Schemas. Therefore, in order to integrate them, they will pass through the XML2RDF component. This component, in conjunction with the ontologies previously mapped from the corresponding XML Schemas, generates the RDF metadata that can be then integrated in the common RDF framework.

This framework has the persistence support of a RDF store, where metadata and ontologies reside. Once all metadata has been put together, the semantic integration can take place, as it is exemplified in section 4.2.

4.2 Semantic Integration Outline

As mentioned in the introduction, one of the main problems in the Wold Wide Web and intraweb domains is that of heterogeneous data integration. Even within a single organization, data from disparate sources must be integrated. Our approach to solve this problem is based on Web ontologies and, as the focus is on multimedia and journalism metadata integration, our integration base are the MPEG-7, MPEG-21 and the journalism ontologies.

In order to profit from the system architecture presented before, when semantic metadata based on different schemes has to be integrated, the XML Schemas are first mapped to OWL. Once this first step has been done, these schemas are easily integrated into the ontological framework using OWL semantic relations for equivalence and inclusion: *subClassOf*, *subPropertyOf*, *equivalentClass*, *equivalentProperty*, *sameIndividualAs*, etc.

These relationships capture the semantics of the data integration. Then, once metadata is incorporated into the system and semantically-decorated, the integration is automatically performed by applying inference. Table 2 shows some of these semantics mappings, performed once all metadata has been moved to the semantic space, plus the preliminary mappings performed during the XML to RDF mapping. The later are necessary in order to recognise implicit identifier, i.e. attributes used to identify instances that are not explicitly used as identifiers in XML. The mappings have been generated using the FOAM ontology mapping framework [28].

Table 2. Journalism and multimedia metadata integration mapping examples

Semantic Mappings
∀ nitf:tobject.subject . subj:Subject
nitf:tobject.subject.detail ≡ subj:explanation
nitf:body ⊆ newsml:DataContent
newsml:Subject ≡ subj:Subject
XML2RDF Mappings
tobject.subject.refnum → rdf:ID

5 Semantic Media Integration from Human Speech

This section introduces a tool, build on top of the ontological infrastructure described in the previous sections, geared towards a convergent and integrated news management in the context of a media house. As it has been previously introduced, the diversification of content in media houses, who must deal in an integrated way with different modalities (text, image, graphics, video, audio, etc.), carries new management challenges. Semantic metadata and ontologies are a key facilitator in order to enable convergent and integrated media management.

In the news domain, news companies like the Diari Segre Media Group are turning into news media houses, owning radio stations and video production companies that produce content not supported by the print medium, but which can be delivered through Internet newspapers. Such new perspectives in the area of digital content call for a revision of mainstream search and retrieval technologies currently oriented to text and based on keywords. The main limitation of mainstream text IR systems is that their ability to represent meanings is based on counting word occurrences, regardless of the relation between words [29]. Most research beyond this limitation has remained in the scope of linguistic [30] or statistic [31] information. On the other end, IR is addressed in the Semantic Web field from a much more formal perspective [32]. In the Semantic Web vision, the search space consists of a totally formalized corpus, where all the information units are unambiguously typed, interrelated, and described by logic axioms in domain ontologies. Such tools enabled the development of semantic-based retrieval technologies that support search by meanings rather than keywords, providing users with more powerful retrieval capabilities to find their way through in increasingly massive search spaces.

Semantic Web based news annotation and retrieval has already been applied in the Diari Segre Media Group in the context of the Neptuno research project [12]. However, this is a partial solution as it just deals with textual content. The objective of the tool described in this section is to show how these techniques can also be applied to content with embedded human-speech tracks. The final result is a tool based on Semantic Web technologies and methodologies that allows managing text and audiovisual content in an integrated and efficient way. Consequently, the integration of human speech processing technologies in the semantic-based approach extends the semantic retrieval capabilities to audio content. The research is being undertaken in the context of the S5T research project[11].

As it is shown in Fig. 3, this tool is based on a human speech recognition process inspired by [33] that generates the corresponding transcripts for the radio and television contents. From this preliminary process, it is possible to profit from the same semi-automatic annotation process in order to generate the semantic annotations for audio, audiovisual and textual content. Keywords detected during speech recognition are mapped to concepts in the ontologies describing the domain covered by audiovisual and textual content, for instance the politics domain for news talking about this subject. Specifically, when the keyword forms of a concept are uttered in a piece of speech, the content is annotated with that concept. Polysemic words and other ambiguities are treated by a set of heuristics. The speech recognition system is first trained with an annotated corpus from news programs performed by different speakers, after which it is ready to be used on the target speech corpus. The speech is parameterized using Mel-frequency cepstral coefficients (MFCC). More details about the annotation and semantic query resolution processes are available from [32].

Once audio and textual contents have been semantically annotated, it is possible to provide a unified set of interfaces, rooted on the semantic capabilities provided by the annotations. These tools, intended for journalists and archivist but also been adapted for the general public, are shown on the left of Fig. 3. They exploit the semantic richness of the underlying ontologies upon which the search system is built. Semantic queries are resolved, using semantic annotations as it has been previously described, and retrieve content items and pieces of these contents. News contents are packaged together using annotations based on the MPEG-21 and MPEG-7 ontologies, as it is described in Section 5.1. Content items are presented to the user through the Media Browser, detailed in Section 5.2, and the underlying semantic annotations and the ontologies used to generate these annotations can be browsed using the Knowledge Browser, described in Section 5.3.

[11] See http://nets.ii.uam.es/s5t

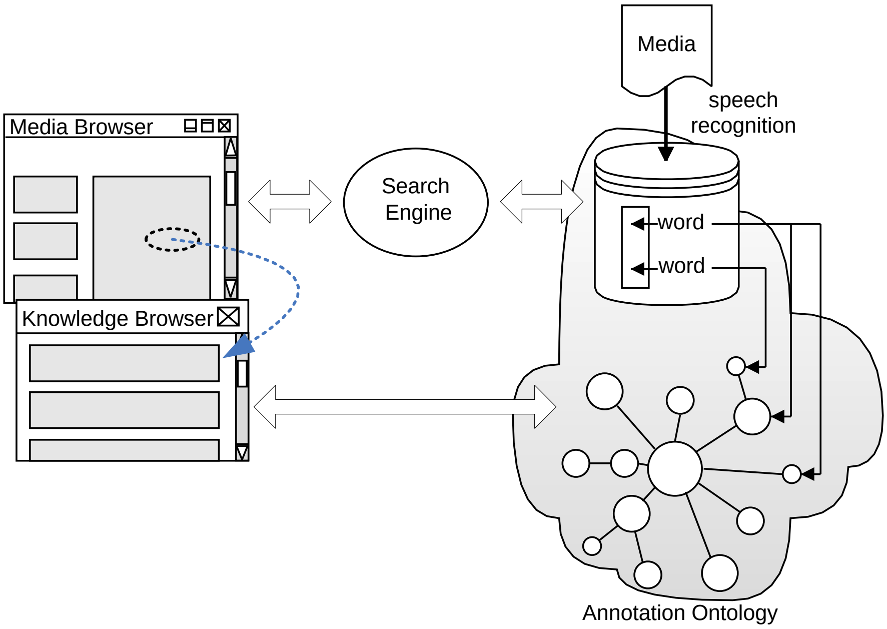


Fig. 3. Architecture for the Semantic Media Integration from Human Speech Tool

5.1 Semantic News Packaging using MPEG Ontologies

Actually, in an editorial office there are a lot of applications producing media in several formats. This is an issue that requires a common structure to facilitate management. The first step is to treat each unit of information, in this case each new, as a single object. Consequently, when searching something upon this structure, all related content is retrieved together.

Another interesting issue is that news can be linked to other news. This link between news allows the creation of information threads. A news composition metadata system has been developed using concepts from the MPEG-21 and MPEG-7 ontologies. It comprises three hierarchical levels as it is shown in Fig. 4.

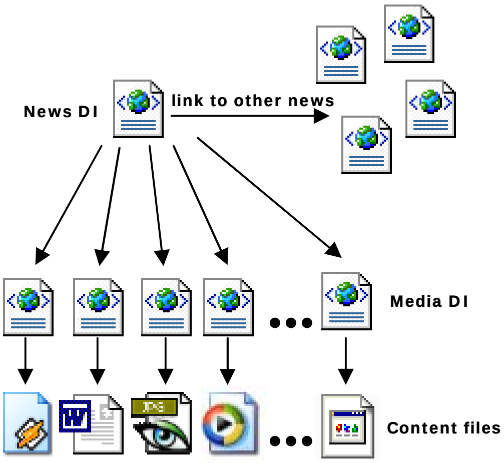


Fig. 4. Content DI structure

The lower level comprises content files, in whatever format they are. The mid level is formed by metadata descriptors (what, when, where, how, who is involved, author, etc.) for each file, mainly based on concepts from the MPEG-7 ontology generated using the methodology described in Section 3.2. They are called the Media Digital Items (Media DI).

These semantic descriptors are based on the MPEG-7 Ontology and facilitate automatised management of the different kinds of content that build up a news item in a convergent media house. For instance, it is possible to generate semantic queries that take profit from the content hierarchy defined in MPEG-7 and formalised in the ontology. This way, it is possible to pose generic queries for any kind of segment (e.g. *AudioSegmentType*, *VideoSegmentType*…) because all of them are formalised as subclasses of *SegmentType* and the implicit semantics can be directly used by a semantic query engine.

Table 3 shows a piece of metadata that describes an audio segment of a Diari Segre Media Group news item used in the S5T project. This semantic metadata is generated from the corresponding XML MPEG-7 metadata using the XML to RDF mapping and takes profit from the MPEG-7 OWL ontology in order to make the MPEG-7 semantics explicit. Therefore, this kind of metadata can be processed using semantic queries independently from the concrete type of segment. Consequently, it is possible to develop applications that process in an integrated and convergent way the different kinds of contents that build up a new.

Table 3. MPEG-7 Ontology description for a news item audio segment generated from XML metadata

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:mpeg7="http://rhizomik.net/ontologies/2006/03/Mpeg7-2001.owl#">
  <mpeg7:AudioType rdf:about="http://rhizomik.net/audio/2007-01-13.mp3">
    <mpeg7:Audio>
      <mpeg7:AudioSegmentType>
        <mpeg7:MediaTime>
          <mpeg7:MediaTimeType>
            <mpeg7:MediaTimePoint
              rdf:datatype="&xsd;time">01:27.0</mpeg7:MediaTimePoint>
            <mpeg7:MediaDuration
              rdf:datatype="&xsd;time">P5S</mpeg7:MediaDuration>
          </mpeg7:MediaTimeType>
        </mpeg7:MediaTime>
      </mpeg7:AudioSegmentType>
    </mpeg7:Audio>
  </mpeg7:AudioType>
</rdf:RDF>
```

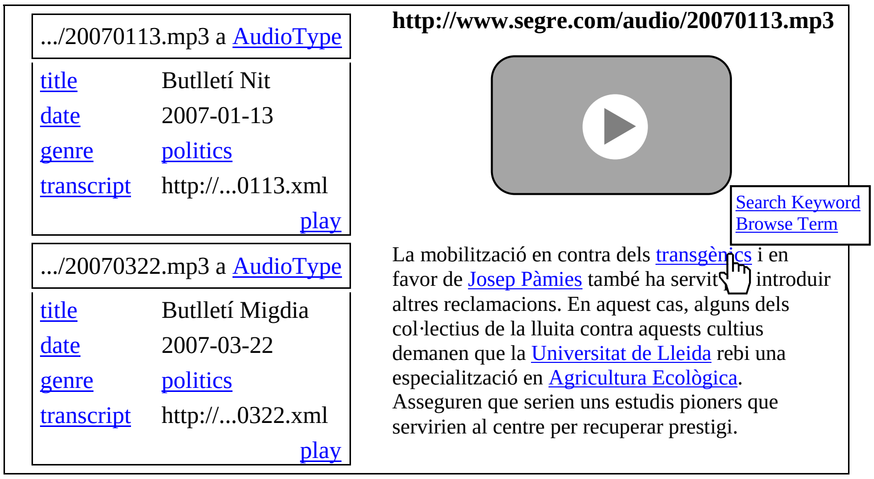
The top level in the hierarchy is based on descriptors that model news and put together all the different pieces of content that conform them. These objects are called News Digital Items (News DI). There is one News DI for each news item and all of them are based on MPEG-21 metadata. The part of the standard that defines digital items (DI) is used for that. DI is the fundamental unit defined in MPEG-21 for content distribution and transaction, very useful for convergent media management. As in the case of MPEG-7 metadata, RDF semantic metadata is generated from XML using the semantics made explicit by the MPEG-21 ontologies. This way, it is possible to implement generic processes also at the news level using semantic queries.

On top of the previous semantic descriptors at the media and news item level, it is possible to develop an application for integrated and convergent news management in the media house. The application is based on two specialised interfaces described in the next subsections. They take profit from the ontological infrastructure detailed in this paper, which is complemented with ontologies for the concrete news domain. However, the application remains independent from the concrete domain.

5.2 Media Browser

The Media Browser takes profit from the MPEG-21 metadata for news and MPEG-7 metadata for media in order to implement a generic browser for the different kinds of media that constitute a news item in a convergent newspaper. This interface allows navigating them and presents the retrieved pieces of content and the available RDF metadata describing them. These descriptions are based on a generic rendering of RDF data as interactive HTML for increased usability [34].

The multimedia metadata is based on the Dublin Core schema for editorial metadata and IPTC News Codes for subjects. For content-based metadata, especially the content decomposition depending on the audio transcript, MPEG-7 metadata is used for media segmentation, as it was shown in Table 3. In addition to the editorial metadata and the segments decomposition, a specialized audiovisual view is presented. This view allows rendering the content, i.e. audio and video, and interacting with audiovisual content through a click-able version of the audio transcript. Fig. 5 shows both the metadata and the audiovisual content views.



.../20070113.mp3 a AudioType

title	Butlletí Nit
date	2007-01-13
genre	politics
transcript	http://...0113.xml

.../20070322.mp3 a AudioType

title	Butlletí Migdia
date	2007-03-22
genre	politics
transcript	http://...0322.xml

La mobilització en contra dels transgènics i en favor de Josep Pàmies també ha servit introduir altres reclamacions. En aquest cas, alguns dels col·lectius de la lluita contra aquests cultius demanen que la Universitat de Lleida rebi una especialització en Agricultura Ecològica. Asseguren que serien uns estudis pioners que servirien al centre per recuperar prestigi.

Fig. 5. The Media Browser interface integrating content metadata and the semantically annotated transcript

Two kinds of interactions are possible from the transcript. First, it is possible to click any word in the transcript that has been indexed in order to perform a keyword-based query for all content in the database where that keyword appears. Second, the transcript is enriched with links to the ontology used for semantic annotation. Each word in the transcript whose meaning is represented by an ontology concept is linked to a description of that concept, which is shown by the Knowledge Browser detailed in the next section. The whole interaction is performed through the user web browser using AJAX in order to improve the interactive capabilities of the interface.

For instance, the transcript includes the name of a politician that has been indexed and modelled in the ontology. Consequently, it can be clicked in order to get all the multimedia content where the name appears or, alternatively, if the name has been semantically annotated, it can be clicked in order to browse all the available knowledge about that politician encoded in the corresponding domain ontology.

5.3 Knowledge Browser

This interface is used to allow the user browsing the knowledge structures employed to annotate content. The same RDF data to interactive HTML rendering used in the Media Browser is used here. Consequently, following the politician example in the previous section, when the user looks for the available knowledge about that person and interactive view of the RDF data modelling him is shown. This way, the user can take profit from the modelling effort and, for instance, be aware of the politician party, that he is a member of the parliament, etc.

This interface constitutes a knowledge browser so the link to the politician party or the parliament can be followed and additional knowledge can be retrieved, for instance a list of all the members of the parliament. In addition to this recursive navigation of all the domain knowledge, at any browsing step, it is also possible to get all the multimedia content annotated using the concept currently being browsed. This step would carry the user back to the Media Browser.

Thanks to this dual browsing experience, the user can navigate through audiovisual content using the Media Browser and through the underlying semantic models using the Knowledge Browser in a complementary an inter-weaved way. Finally, as for the Media Browser, the Knowledge Browser is also implemented using AJAX so the whole interactive experience can be enjoyed using a web browser.

6 Conclusions and future work

This research work has been guided by the need for a semantic journalism and multimedia metadata framework that facilitates semantic newspaper applications development in the context of a convergent media house. It has been detected, as it is widely documented in the bibliography and professional activity, that IPTC and MPEG standards are the best sources for an ontological framework that facilitates a smooth transition from legacy to semantic web era systems. MPEG-7, MPEG-21 and most of the IPTC standards are based on XML Schemas and thus they do not have formal semantics.

Our approach contributes a complete and automatic mapping of the whole MPEG-7 standard to OWL, of the media packaging part of MPEG-21 and of the main IPTC standard schemas (NITF, NewsML,...) to the corresponding OWL ontologies. Instance metadata is automatically imported from legacy systems through a XML2RDF mapping, based on the ontologies previously mapped from the standard XML schemas. Once in a semantic space, data integration, which is a crucial factor when several sources of information are available, is facilitated enormously.

Therefore, though the semantics that can be reused from XML Schemas are lightweight, they provide the anchor points where more elaborated formalisations can be connected. For instance, the COMM Ontology [35], which formalises parts of MPEG-7, might be integrated with the proposed MPEG-7 ontology. This way, more advanced reasoning capabilities might be available and initiatives like the COMM Ontology might enjoy input metadata coming from MPEG-7 XML data in order to enlarge the scenarios where they can be put into practice. Another initiative we are currently considering is NewsML G2[12]. Like NewsML, it is also based on XML Schemas so we are going to apply the same approach to it.

Finally, semantic metadata facilitates the development of applications in the context of media houses that traditional newspapers are becoming. The convergence of different kinds of media, that now constitute multimedia news, poses new management requirements that are easier to cope with if applications are more informed, i.e. aware of the semantics that are implicit in news and the media that constitute them. The benefits of semantic metadata are being tested in the Diari Segre Media Group, a newspaper that is becoming a convergent media house with press, radio, television and an Internet portal. As it has been detailed, a set of semantics-aware tools have been developed. They are intended for journalist and archivists in the media house, but they can be also adapted to the general public at the Internet portal.

Acknowledgements

This work has been partly supported by the Spanish Ministry of Science and Education research project S5T - Scaleable semantic personalised search of spoken and written contents on the Semantic Web (TIN2005-06885) http://nets.ii.uam.es/s5t

References

1. Eriksen, L. B. and Ihlström, C., Evolution of the Web News Genre – The Slow Move Beyond the Print Metaphor, in: Proceedings of the 33rd Hawaii international Conference on System Sciences, IEEE Computer Society, 2000.
2. Lundberg, J.. The online news genre: Visions and state of the art, presented at the 34th Annual Congress of the Nordic Ergonomics Society, Sweden, 2002.
3. Ihlström, C., Lundberg, J. and Perdrix, F., Audience of Local Online Newspapers in Sweden, Slovakia and Spain - a comparative study, in: Proceedings of HCI International, 2003.
4. McDonald, N., Can HCI shape the future of mass communications? Interactions 11(2) (2004) 44-47.
5. Berners-Lee, T., Hendler, J., Lassila, O., The Semantic Web. Scientific American, 2001.
6. Doerr, M., Hunter, J. and Lagoze, C., Towards a Core Ontology for Information Integration, Journal of Digital Information 4(1) (2003).
7. Troncy, R., Integrating Structure and Semantics into Audio-visual Documents, in:. Proceedings of the 2nd International Semantic Web Conference, Florida, USA, 2003.
8. Hunter, J., Enhacing the Semantic Interoperability of Multimedia through a Core Ontology, IEEE Trans. on Circuits and Systems for Video Technology 13(1) (2003) 49-58.
9. Tsinaraki, C., Polydoros, P. and Christodoulakis S., Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing, in: Proceedings of the 16th International Conference on Advanced Information Systems Engineering, 2004.

[12] NewsML G2 Architecture (NAR), http://www.iptc.org/NAR

10. Salembier, P. and Smith, J., Overview of MPEG-7 multimedia description schemes and schema tools, in: Manjunath, B.S.; Salembier, P. and Sikora, T. (ed.), Introduction to MPEG-7: Multimedia Content Description Interface, John Wiley and Sons, 2002.

11. Haustein, S. and Pleumann, J., Is Participation in the Semantic Web too Difficult? in: Proceedings of the International Semantic Web Conference. Sardinia, Italy, 2002.

12. Castells, P., Perdrix, F., Pulido, E., Rico, M., Benjamins, R., Contreras, J. and Lorés, J., Neptuno: Semantic Web Technologies for a Digital Newspaper Archive, Springer, LNCS 3053, 2004, pp. 445-458.

13. Fernández-García, N. and Sánchez-Fernández, L., Building an Ontology for NEWS Applications, Poster at the International Semantic Web Conference, 2004.

14. Hunter, J., A Proposal for an MPEG-7 Description Definition Language, MPEG-7 AHG Test and Evaluation Meeting, Lancaster, 1999.

15. Hunter, J., Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology, International Semantic Web Working Symposium (SWWS), 2001.

16. García, R., Gil, R. and Delgado, J., A web ontologies framework for digital rights management, Journal of Artificial Intelligence and Law 15(2) (2007) 137-154.

17. García, R., Gil, R., Gallego, I. and Delgado, J., Formalising ODRL Semantics using Web Ontologies, Open Digital Rights Language Workshop, ODRL'05, 2005.

18. García, R. and Celma, O., Semantic Integration and Retrieval of Multimedia Metadata, Knowledge Mark-up and Semantic Annotation Workshop, Semannot'05, CEUR Workshop Proceedings 185 (2005) 69-80.

19. Klein, M.C.A., Interpreting XML Documents via an RDF Schema Ontology, in: Proceedings of the 13th Int. Workshop on Database and Expert Systems Applications, pp. 889-894, 2002.

20. Amann, B., Beeri, C., Fundulaki, I. and Scholl, M., Ontology-Based Integration of XML Web Resources, in: Proceedings of the 1st International Semantic Web Conference, pp. 117-131, 2002.

21. Cruz, I., Xiao, H. and Hsu, F., An Ontology-based Framework for XML Semantic Integration, in: Proceedings of the 8th Int. Database Engineering and Applications Symposium, Portugal, 2004.

22. Lakshmanan, L. and Sadri, F., Interoperability on XML Data, in: Proceedings of the 2nd International Semantic Web Conference, 2003.

23. Patel-Schneider, P.F. and Simeon, J., The Yin/Yang web: XML syntax and RDF semantics, in: Proceedings of the 11th World Wide Web Conference, pp. 443-453, 2002.

24. García, R., XML Semantics Reuse, Chapter 7 in: A Semantic Web Approach to Digital Rights Management, PhD Thesis, Universitat Pompeu Fabra, Barcelona, 2006. http://rhizomik.net/~roberto/thesis

25. Berners-Lee, T., Why RDF model is different from the XML model, W3C Dessign Issuer, September 1998. http://www.w3.org/DesignIssues/RDF-XML.html

26. Tous, R., García, R., Rodríguez, E. and Delgado, J., Arquitecture of a Semantic XPath Processor, in: Bauknecht, K., Pröll, B. and Werthner, H. (eds.) E-Commerce and Web Technologies, EC-Web’05, Springer, LNCS 3590, pp. 1-10, 2005.

27. Mark van Assem et al., A Method to Convert Thesauri to SKOS, in: Proceedings of the Third European Semantic Web Conference (ESWC'06). Springer, LNCS 4011, pp. 95-109, 2006.

28. Ehrig, M. and Sure, Y., FOAM - Framework for Ontology Alignment and Mapping; Results of the Ontology Alignment Initiative, in: B. Ashpole, M. Ehrig, J. Euzenat, H. Stuckenschmidt (eds.) Proceedings of the Workshop on Integrating Ontologies, CEUR Workshop Proceedings 156 (2005) 72-76.

29. Salton, G.,and McGill, M., Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

30. Vorhees, E., Query expansion using lexical semantic relations, in Proceedings of the 17th ACM Conf. on Research and Development in Information Retrieval, Dublin, Ireland, 1994.

31. Cuayahuitl, H. and Serridge, B., Out-of-vocabulary Word Modelling and Rejection for Spanish Keyword Spotting Systems, in: Proceedings of the 2nd Mexican International Conference on Artificial Intelligence (MICAI 2002), Mérida, Mexico, 2002.

32. Castells, P,; Fernández, M. and Vallet, D., An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval, IEEE Transactions on Knowledge and Data Engineering 19(2) (2007) 261-272.

33. Kim, J. et al., A Keyword Spotting Approach based on Pseudo N-gram Language Model, in: Proceedings of the 9th Conf. on Speech and Computer (SPECOM 2004), Patras, Greece, 2004.

34. García, R. and Gil, R., Improving Human–Semantic Web Interaction: The Rhizomer Experience, in: Proceedings of the 3rd Italian Semantic Web Workshop (SWAP'06), CEUR Workshop Proceedings 201 (2006) 57-64.

35. Arndt, R., Troncy, R., Staab, S., Hardman, L.and Vacura, M., COMM: Designing a Well-Founded Multimedia Ontology for the Web, in: Proceedings of the 6th International Semantic Web Conference (ISWC'2007), Busan, Korea, 2007.