

Newspaper Archives on the Semantic Web

**P. Castells¹, F. Perdrix², E. Pulido¹, M. Rico¹, J. M. Fuentes¹,
V. R. Benjamins³, J. Contreras³, E. Piqué², J. Cal², J. Lorés⁴,
T. Granollers⁴**

¹Universidad Autónoma de Madrid

Campus de Cantoblanco, C/ Tomás y Valiente 11, 28049 Madrid
{pablo.castells,estrella.pulido,mariano.rico,chema.fuentes}@uam.es

²Diari Segre S.L.U.

C/ Del Riu nº6, 25007 Lleida

{fperdrix,epique,jcal}@diarisegre.com

³iSOCO, S.A.

C/ Fca. Delgado 11 – 2º 28100 Alcobendas – Madrid

{rbenjamins,jcontreras}@isoco.com

⁴Universitat de Lleida

C/ Jaume II nº69, 25001 Lleida

{jesus,tonig}@griho.net

1 Introduction

The introduction of information technologies in the news industry has marked a new evolutionary cycle in the journalistic activity. The creation of new infrastructures, protocols and exchange standards for the automatic or on-demand distribution and/or sale of information packages through different channels and transmission formats has deeply transformed the way in which news industry players communicate with each other. One interesting consequence of this technological transformation has been the emergence, in very few years, of a whole new market of online services for archive news redistribution, syndication, aggregation, and brokering. Newspaper archives are a highly valuable information asset for the widest range of information consumer profiles: students, researchers, historians, business professionals, the general public, and not the least, news writers themselves. Providing technology for news archive construction, management, access, publication, and billing, is an important business nowadays.

The information collected from everyday news is huge in volume, very loosely organized, and grows without a global a-priori structure. This ever-growing corpus of archived news results from the coordinated but to much extent autonomous work of a team of reporters, whose primary goal is not to build an archive, but to serve the best possible information product for

immediate consumption. Reporters are often assisted by librarians and archive specialists, who help classify, index, and annotate news as they are sent to the archive, using special-purpose archive management software. But in addition to this, powerful search and navigation mechanisms are needed for information consumers to find their way through. Current technology typically provides keyword-based search, browsing facilities inside newspaper issues, and, in online newspapers, navigation through static hand-made hyperlinks between news materials.

A wide margin remains yet for taking advantage of the possibilities offered by the digital medium to exploit a newspaper archive. Aspects that can be improved include: a) keyword search falling short in expressive power; b) weak interrelation between archive items: users may need to combine several indirect queries manually before they can get answers to complex queries; c) lack of a commonly adopted standard representation for sharing archive news across newspapers; d) lack of internal consensus for content description terminology between and among reporters and archivists; e) lack of involvement of reporters in the archiving process. We believe the emerging Semantic Web technologies [1] provide a good approach to overcome these limitations.

The Neptuno project (see <http://nets.ii.uam.es/neptuno>) has been set up to apply Semantic Web technologies to improve current state of the art in diverse aspects of the production and consumption of digital news. This paper presents the results achieved in the first phase of the project, which focuses on the construction, management and exploitation of a newspaper archive. The goal of this work is to develop a high-quality semantic archive for the Diari SEGRE newspaper where a) reporters and archivists have more expressive means to describe and annotate news materials, b) reporters and readers are provided with better search and browsing capabilities than those currently available, and c) the archive system is open to integration in potential electronic marketplaces of news products.

According to these goals, a platform has been developed whose main components are:

- An ontology for archive news, based on journalists' and archivists' expertise and practice, and integrating current dominant standards from the IPTC consortium [2].
- A knowledge base where archive materials are described using the ontology. A DB-to-ontology conversion module automatically integrates existing legacy archive materials into the knowledge base.
- A semantic search module, where meaningful information needs can be expressed in terms of the ontology, and more accurate answers are supplied.

- A visualization and navigation module to display individual archive items, parts or combinations of items, and lists or groups of items.

The Diari SEGRE reporters will be the primary users of the archive exploitation functionalities. A version for the general public is planned as a future extension of the project.

The rest of the paper is organized as follows. The next section describes the creation and management of the newspaper library with the technology previously in use at the Diari SEGRE. After this, an overview of the online newspaper archive industry and current technologies is given. Section 3 describes the definition of an ontology for the Neptuno project, and Section 4 explains the search and visualization functionalities for the knowledge base. Section 5 describes the platform architecture and the construction of the knowledge base from current archive news databases, providing some implementation details.

2 Online Newspaper Archives

The elaboration of news within a mass media group like Diari SEGRE is fed by diverse information sources, among which in-house newspaper archives are a must. These digital archives are a constant reference for background information search, or browsing related news, in order to complement, clarify, or help place in a certain context the new information a journalist is writing.

Most press media nowadays have massive information repositories based on relational database systems, with special-purpose software to manage their contents. The Diari SEGRE newspaper may publish over a hundred news and some fifty photographs everyday, which are automatically uploaded into the digital archive of the newspaper. The archive consists of a database, JPEG image files (published photos), and PDF files (newspaper pages). The archive currently contains all the issues of the newspaper since July 1995 to date. Milenium Arcano, one of the most popular content management platforms in the news sector in Spain and Portugal, is used to manage the archive. Arcano provides functionalities for archive update, manual documentation, and content search.

The news archival process is done by the documentation department, according to the criteria of experts in this department, and the possibilities (and limitations) of the software platform, classifying news using a hierarchical thesaurus, in endless evolution, of available concepts for contents annotation. On the other side, journalists consume this archived information when they need to inform themselves on subjects, histories or events,

often under strong time constraints, and with limited beforehand archive system knowledge to be sure how to formulate their queries and information needs.

Today virtually all press media published on the web have a public access system to their news archive through the web. While a few newspapers keep providing free archive access, in the last few years paid access, by pay-per-view or periodical subscription, is becoming prevalent. Diari SEGRE itself has a service of this kind. Online archives usually provide searching and browsing facilities for archive news, which, in their advanced modality, allow queries by section, date, heading, lead, body, author, type of content (text, photography, graphic, animation, audio, video), etc. These applications suffer, at a smaller scale, from the same problems and limitations as the ones highlighted by the Semantic Web perspective: no support for conceptual search; extensive ad-hoc implementation efforts are required for integration with other archives or external information systems; platforms are not open to unforeseen extensions; rigid browsing facilities; no explicit notion of the semantics conveyed by archive documents. The aim of our project is to achieve or enable specific improvements by introducing ontology-based semantics, and exploiting this to provide better and/or novel functionalities in a news archive management system, following and/or improving existing proposals from the Semantic Web field, and contributing our own.

3 An Ontology for a Newspaper Library

The first step in the development of the Neptuno project has been the definition of an ontology to represent and process news information. After evaluating the available languages and standards for ontology definition, we have chosen RDF [3], currently the most mature, stable and widespread standard in the latest projects and developments in the Semantic Web area.

The level of detail of domain semantics in our ontology stops at the definition of generic categories by topics and subtopics, such as “politics”, “immigration”, “economy”, “trade”, “stock market”, or “sports”, as will be described next, but does not include specific classes and entities for these areas, such as “political party”, “suffrage”, “judge”, “lawyer”, “sentence”, “sportsman”, “theatre play”, or “music group”, nor the instances of these entity types. The creation of a knowledge base for these entities, that would completely take in the potential informative coverage of a mass media, exceeds the capacity of any single organisation willing to undertake

such an endeavor. On the other hand, the utility of a partial collection, more feasible to construct, is difficult to justify.

Our first observation, when considering the reuse of existing ontologies and standards in this field, is that as of now no proper journalistic ontology has been published, as far as we are aware. In this sense our work is a contribution to the growth of the Semantic Web and publicly available ontology collections. With respect to other kind of controlled vocabularies, different standards have been developed in the area of journalism, such as NewsML [4], NITF [5], XMLNews [6], the IPTC subject reference system [2], and PRISM [7]. After evaluating all these standards, we have adopted the IPTC Subject Reference System, transforming it into an RDFS class hierarchy, as a thematic classification system for news archive contents.

Diari SEGRE has a database in which information about news, photographs, graphics and pages is stored. Newspaper contents are classified everyday by archivists using the Milenium Arcano tool. The criteria used for this classification are basically two: the section to which the contents belongs (Sports, Economy, ...) and the topics they deal with. For the latter classification a thesaurus is used that has been elaborated incrementally over time among all archivists, and that is frequently updated as new needs arise.

The ontology to represent the archives of Diari SEGRE has been built by using the Protégé ontology editor [8]. Some of the concepts in this ontology correspond to tables in the existing database, such as News, Photograph, Graphics and Page. All these concepts are subclasses of the Contents concept. In addition, contents are classified by subject following the IPTC subject classification hierarchy. This classification replaces those by section and category in the thesauri that were used so far by archivists. A mapping has been established between the old thesaurus categories and those in the IPTC hierarchy.

4 Semantic Query and Semantic Navigation

The defined ontology serves to enrich the existing news archive contents with explicit semantic representations, giving rise to a semantic knowledge base. The added value of this enrichment pays off with the possibility to develop advanced exploitation modules like, in Neptuno, a semantic query module, and a system for ontology visualization, both integrated in a semantic portal.

4.1 Semantic Query

With respect to search, the availability of semantic information in the knowledge base allows the user to formulate more precise and expressive user queries, and implement a system that is able to use conceptual elements to match information needs against archive contents. Most existing search systems today are keyword-based: the user introduces the relevant words, and the search engine retrieves all the documents that contain them. Occurrences of words are sought in documents, without taking into account:

- The meaning of words (they may have multiple).
- The relation between words in the query.

Which can result in the following problems:

- The system may return many documents with low relevance for the original query.
- It is the user's responsibility to open each document to check its relevance.

This may result in users not finding the sought information even when it exists, or on the other extreme, an information overload due to the large number of documents returned.

A semantic search engine [9] has knowledge of the domain at hand. The availability of a domain ontology that structures and relates the information according to its meaning allows the implementation of a search system where users can specify search criteria in terms of modeled concepts and attributes. The results are presented in a structured form including only the requested information. Contrary to traditional search systems where the answer consists of whole documents, where the user has to find manually the sought information, semantic search systems can return only the requested information.

The search module in Neptuno has been developed following these principles of semantic search. Moreover, the module combines direct search by content classes and class fields, with the possibility to browse the IPTC taxonomy, according to which archive news and documents are classified. The combination is twofold. On the one hand, search by fields is restricted to the IPTC categories selected by the user. On the other, Neptuno shows the list of categories to which the results of a search belong. The user can navigate directly to them, or select them to narrow down the search successively.

4.2 Visualization

Despite the advantages that semantic models provide to retrieve information, one of the problems of these models is supplying a readable and understandable presentation for the end-user. In the model design and construction phase, the expressive value of the model is valued, and no visual or aesthetic aspects are taken into consideration. The main purposes for building ontologies are to provide semantic content for intelligent systems. The knowledge models are designed to offer the appropriate information to be exploited by the software. No visualization criteria are used to build an ontology and often the information is not suitable to be published as it is. For example, it may happen that concepts have too many attributes, or relations are represented as first class objects so that the navigation becomes tedious, or concepts to be shown do not always correspond to modelled ones.

In our case, modelling of concepts and relations in the newspaper archive ontology has not been restricted by publication criteria, as could be the number of attributes of a concept, number of instances, or existence of auxiliary concepts for the representation of relations. There is a need to differentiate between what is going to be modelled from how it is going to be visualized. That is why we introduce the concept of Visualization Ontology [10]. This ontology, called publication schema, allows organising the concepts and attributes in order to be published in the portal. The visualization ontology represents publication concepts as they should appear in the portal. It does not duplicate the content of the original ontology, but links the content to publication entities using an ontology query language. This way one ontology that represents a particular domain can be visualized through different views. The visualization ontology has two predefined concepts:

- **Publication Entity:** encapsulates objects as they will be published. Any concept defined in the visualization ontology will inherit from it and should define these attributes:
 - XSL style-sheet associated to the concept that translates its instances to HTML.
 - Query that retrieves all attribute values from the original ontology.
- **Publication Slot:** Each attribute that is going to appear on the web should inherit from this concept. Different facets describe how the attribute will appear on the page.
 - **Web label:** The label that will appear with the value.
 - **RDQL:** reference to the query used to retrieve the attribute value.
 - **Link:** When the published value should perform some action on mouse click (link, email, button, etc.), the action is described here.

Portal elements are described as children of the Publication Entity and their instances are defined according to the languages the entity will be published in (labels in English, Spanish, etc.), or the channel (the transformation style-sheet can translate into HTML, WAP, or just XML). In this case, the news library is exported in HTML format.

5 Architecture and Implementation

The current version of the Neptuno platform is an extension of previously working systems at Diari SEGRE that does not interfere with technology and procedures previously in use. The manual creation of ontology instances, associated to several hundred thousands news accumulated over the last nine years in the Diari SEGRE archive, would unquestionably be the best way to warrant the highest quality of the introduced semantics, but also an out of proportion, unfeasible work, out of any reasonable cost/benefit balance. For this reason Neptuno includes a module that populates the ontology with instances that are extracted automatically from the Arcano database, by using a mapping tool from JDBC/ODBC databases to RDF [11].

The manual creation of instances for new information and photographs stored in the newspaper archive, at a regular daily pace, from the deployment of Neptuno onwards, 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. Meanwhile, we have decided not to interfere with the current environments with which the Diari workers interact, and to generate new instances everyday by the same automatic procedure as used for old materials, without altering the newspaper production pace.

The contents search and visualization modules operate directly on the knowledge base in RDF. The user poses search requests through a web interface in which (s)he selects the contents class to be searched (News, Photograph, Graphics, or Page), and specifies keywords for the desired fields (heading, author, section, date, subject, etc.) in the selected class. This information is sent to the Neptuno server where the request is formalized as an RDQL query. This query is run on the knowledge base and returns a list of resources (instances) that satisfy the query constraints.

The list of resources is sent to the visualization module, which generates an web page where the found instances are displayed in an abbreviated and clickable form. When the user selects an instance, the visualization module

finds the class to which the instance belongs, selects the corresponding publishing entity, and generates a web page where the resource details are shown according to the principles described in the previous section.

6 Conclusions and Future Work

A newspaper archive is a fundamental working tool for editorial teams, and a potentially marketable product towards different kinds of consumers through diverse distribution channels. The size and complexity of the stored 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.

The work developed so far represents an application experience of Semantic Web technologies in a real setting, making novel contributions to several of the undertaken aspects: definition of ontologies in a specific domain, semantic search and exploration functionalities, development of a user interface to interact with the knowledge base, transition from a working system with traditional technologies to a semantic-based platform.

Other goals we are considering from this point are the integration of new sources (external newspaper archives), contents types (text, audio, video) and languages (Spanish and Catalan), the enrichment of contents description through automatic methods of text and multimedia analysis, and the automatic adaptation to multiple devices and access channels. We also intend to improve the achievements so far by, for example, increasing the precision of the search system by defining ranking algorithms, increasing its expressiveness with more advanced interfaces, and carrying out a methodological revision of the representation and classification systems (ontology) currently used in the knowledge base.

Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology, grants FIT-150500-2003-511 and TIC2002-1948.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* (2001).
2. IPTC Subject Reference System & NewsML Topicsets, <http://www.iptc.org/metadata>
3. Lassila, O., Swick, R.R.: Resource Description Framework (RDF) Model and Syntax Specification. W3C Rec. 22 Feb. 99. Available at <http://www.w3.org/TR/REC-rdf-syntax>
4. IPTC NewsML, <http://www.newsml.org>
5. IPTC News Industry Text Format (NITF), A Solution for Sharing News, <http://www.nitf.org>
6. XMLNews, XML and the News Industry, <http://www.xmlnews.org>
7. Pub. Requirements for Industry Standard Metadata (PRISM), <http://www.prismstandard.org>
8. Noy, N.F., Sintek, M., Decker, Crubezy, M., Ferguson, R.W., Musen, M.A.: Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems* 16(2) (2001) 60-71.
9. Guha, R., McCool, R., Miller, E.: Semantic search. *12th International World Wide Web Conference (WWW 2003)*, Budapest, Hungary (2003), 700-709
10. Contreras, J., Benjamins, V.R., Prieto, J.A., Patón, D., Losada, S., González, D.: Duontology: an Approach to Semantic Portals based on a Domain and Visualisation Ontology. *KTWeb*.
11. Bizer, C.: D2R MAP - A Database to RDF Mapping Language. *12th International World Wide Web Conference (WWW 2003)*. Budapest, Hungary (2003).