Neptuno: Semantic Web Technologies for a Digital Newspaper Archive

P. Castells[1], F. Perdrix[2], E. Pulido[1], M. Rico[1],
R. Benjamins[3], J. Contreras[3], J. Lorés[4]

[1]Universidad Autónoma de Madrid
Ctra. de Colmenar Viejo km. 15, 28049 Madrid
{pablo.castells,estrella.pulido,
mariano.rico}@uam.es
[2]Diari Segre S.L.U.
C/ Del Riu nº6, 25007 Lleida
fperdrix@diarisegre.com
[3]iSOCO, S.A.
c/ Fca. Delgado 11 – 2º 28100 Alcobendas – Madrid
{rbenjamins,jcontreras}@isoco.com
[4]Universitat de Lleida
C/ Jaume II nº69, 25001 Lleida
jesus@griho.net

Abstract. Newspaper archives are a fundamental working tool for editorial teams. Their exploitation in digital format through the web, and the provision of technology to make this possible, are also important businesses today. The volume of archive contents, and the complexity of human teams that create and maintain them, give rise to diverse management difficulties. We propose the introduction of the emergent semantic-based technologies to improve the processes of creation, maintenance, and exploitation of the digital archive of a newspaper. We describe a platform based on these technologies, that consists of a) a knowledge base associated to the newspaper archive, based on an ontology for the description of journalistic information, b) a semantic search module, and c) a module for content browsing and visualisation based on ontologies.

1 Introduction

The introduction of information technologies in the news industry has marked a new evolutionary cycle in journalistic activity. Digital media allow an unprecedented dissemination, ease of access, immediacy, economy, virtually unlimited extension, and *à la carte* information delivery, eliminating time and space restrictions, and altering editorial team routines. The creation of new infrastructures, protocols and exchange standards for

the automatic (push) or on-demand (pull) distribution and/or sale of information packages through different channels and transmission formats has deeply transformed the way in which the different specialised agents that participate in the news industry (companies, media, groups, agencies, consortiums, professionals, etc.) communicate with each other. Internally, the trend for information producers points to the adoption of integrated platforms that support the whole cycle of contents elaboration, management, and publication, spanning from the reception of external information (e.g. from news agencies), the elaboration of own contents, layout composition, documentation, archive management, etc.

One interesting consequence of this technological transformation in the media industry has been the emergence, in very few years, of a whole new market of online services for archive news redistribution, syndication, aggregation, and brokering (see for example NewsLibrary [1], the British Library [2], or a list of online U.S. newspaper archives [3]). 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 (see for instance NewsViews Solutions [4], ActivePaper Archive [5]).

The information collected from everyday news is huge in volume (e.g. by mid 2003 LexisNexis [6] claimed to handle over 3.3 billion documents), very loosely organised (e.g. compared to a book library), and grows without a global a-priori structure, as news stories add up and evolve unpredictably. 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.

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 (often by fields: body, headline, section, lead, byline), browsing facilities inside newspaper issues, and, in online newspapers, navigation through static hand-made hyperlinks between news materials (e.g. links to earlier background stories).

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 [7] provide a good approach to overcome these limitations.

The Neptuno project[1] 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. It is being conducted by two universities (Universidad Autónoma de Madrid and Universitat de Lleida), a news media company (Diari SEGRE), and a technology provider (iSOCO, S. A.). 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 the work described here 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 [8].
- 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 visualisation 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 organised 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 4 describes the definition of an ontology for the Neptuno project, and Section 5 explains the search and visualisation functionalities for the knowledge base. Section 6 describes the platform architecture and the construction of the knowledge base from current archive news databases, providing some implementation details.

2 Document Management at Diari SEGRE

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

[1] http://seweb.ii.uam.es/neptuno/

news, in order to complement, clarify, or help place in a certain context the new information a journalist is writing.

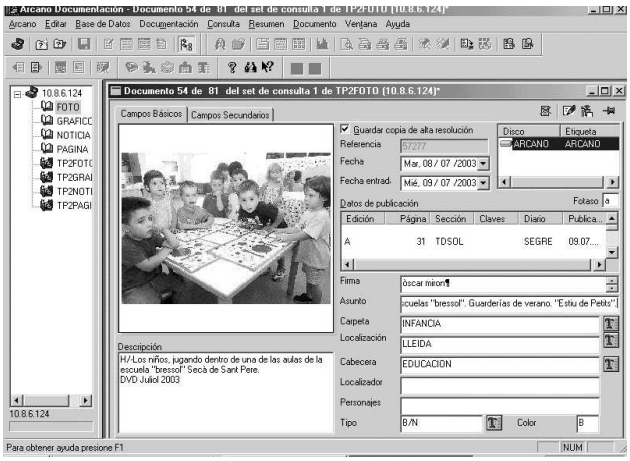

Fig. 1. Current news archive management environment

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. This means nearly 400,000 news and 200,000 photos, which take about 14 Gb total disk space. A software platform named Milenium Arcano [9] (see figure 1), one of the most popular ones 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 previously stored information when they need to inform themselves on subjects, histories or events, not always with the ideally desirable available time, neither with enough beforehand archive system knowledge to be sure how to formulate their queries and information needs. Because of all this, the differences between archivists' and journalists' mental models (evidenced along this project) call for a more flexible content categorisation and search system, which is the aim of the Neptuno project, by making use of Semantic Web technologies.

3 Online Newspaper Archives on the WWW

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. The Special Libraries Association News Division (SLAND) maintains a list of online newspaper archives at http://www.ibiblio.org/slanews/internet/archives.html.

Online archives usually provide searching and browsing facilities for archive news. For instance, the search service of the newspaper El País [10] (the one with most readers in Spain), in its most advanced modality, allows queries by section, date, heading, lead, body, author, and type of content (text, photography, graphic, animation, audio, video). This kind of search options, with very slight variations, is representative of thousand other archive services offered on the WWW by media companies from the five continents.

Other more ambitious projects integrate the archives of several newspapers, and even recover the old historic issues conserved in print. NewsLibrary [1], for example, offers online access to more that 200 newspapers and other sources (covering different time spans, the oldest ones starting in 1977), including all first-rank news media. The British Library [2] provides online access to more that half a million articles from five historic newspapers (Daily News, News of the World, Penny Illustrated, The Manchester Guardian, Weekly Dispatch), including conserved issues from as far back as 1851, properly digitized, segmented and graphically treated. ProQuest Historical Newspapers [11] is a similar product that integrates one and a half century of contents (news, advertisements, comic strips, letters to the editor, weather information, and other content genres) of The New York Times (1851-2001), The Wall Street Journal (1889-1987), The Washington Post (1877-1988), The Christian Science Monitor (1908-1991), and Los Angeles Times (1881-1984). These applications allow searching and navigating through old issues, and visualising the pages in HTML or the original format.

With the emergence of this new market, technology and solution providers for the deployment of online archives have proliferated as well, addressing all aspects needed for the development of the application, depending on the starting materials (paper, microfiche, digital), the target environment (web, intranets), and the intended use (internal, public, commercial). This may involve digitation and graphical treatment of materials, archive management and maintenance, integration in the information production chain, publication, access, etc. Many commercial tools exist today to this end, like ActivePaper Archive [5] from OliveSoftware, NewsViews Solutions tools [4], ArchiveIQue [12] from Baseview Products, Canto Cumulus [13], or DC4 [14] from Digital Collections, to name just a few.

These and similar available platforms and 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.

4 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 [15], currently the most mature, stable and widespread standard in the latest projects and developments in the Semantic Web area.

According to the reference methodologies in the Semantic Web, the recommended steps for the construction of an ontology are [16]:

1. Determine the domain and scope of the ontology.
2. Determine the intended use of the ontology.
3. Reuse existing ontologies or controlled vocabularies.
4. Enumerate important terms in the domain.
5. Define the class hierarchy.
6. Create instances.

In the Neptuno project we have taken this recommendation as a general guide, adapting it to our particular case without significant deviations. We describe next the definition process for each of these aspects, and the resulting ontology. The last of the above-mentioned steps, instance creation, which is done automatically in Neptuno, is described later in Section 6 about platform architecture.

4.1 Domain, Scope and Intended Use

Although, apparently, the selection of the domain should be an obvious question, the journalistic field has the peculiarity of potentially dealing with topics in all fields of human knowledge and current affairs: politics, culture, courts, science, sports, art, economy, etc. It has been necessary to carefully establish a limit in the domains to represent, without which any attempt at approaching completitude would lead to a whole project for each thematic information area.

The conceptual reflection of these thematic areas in our ontology is limited to 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", "actress", "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, company or agency that would intend to undertake such an endeavor. On the other hand, the utility of a partial collection, more feasible to construct, is difficult to justify.

As for the intended use, after a thorough analysis, we have come to the conclusion that the management of a newspaper archive is, of all the potentially targetable aspects in the production cycle of a newspaper, possibly the one which offers more and best opportunities for improving processes and products, and best lends itself to the Semantic Web proposals. Among other reasons that sustain this assessment, we can mention a) the quality of the newspaper library contents, which have passed a selection filter (as opposed to the news flow that arrive everyday to the newspaper offices), given that only the news that actually make it to the newspaper pages are stored in the archive; b) the enrichment of contents with metadata, descriptions, and a careful manual categorisation by professional archivists who supervise the transit of news to the archive one by one; and c) the persistence of materials for an indefinite period of time (once again, in contrast to more ephemeral materials from agencies and other sources, that are finally discarded), which allows a continued exploitation of the added value that results from the contributions we propose here.

4.2 Existing Ontologies and Controlled Vocabularies

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 [17], NITF [18], XMLNews [19], the IPTC subject reference system [8], and PRISM [20]. NewsML and NITF (News Industry Text Format) are XML-based standards to represent and manage news along their whole lifecycle, including their creation, exchange and consumption. While NewsML is used to represent news as multimedia packages, NITF deals with document structure. XMLNews is a subset of NITF and is based on RDF. It includes a set of tags (such as location, person, or date) that allow annotating news to facilitate information search.

These three standards have been created by the IPTC (International Press Telecommunications Council), an international consortium of news agencies, editors and newspapers distributors. This organism has proposed the Subject Reference System, a subject classification hierarchy with three levels and seventeen categories in its first level.

PRISM (Publishing Requirements for Industry Standard Metadata) was developed by IDEAlliance (International Digital Enterprise Alliance), an industrial board of editorial companies and publishing software manufacturers which includes Adobe, Quark, Condé Nast and Time Inc. It is an XML standard, similar to NewsML, that provides a metadata vocabulary for the editorial industry to facilitate aggregation and syndication of digital contents.

After evaluating all these standards, we have adopted the IPTC Subject Reference System as a thematic classification system for news archive contents. The integration and adaptation of this standard to our ontology has been carried through by a) converting the IPTC topic hierarchy to an RDF class hierarchy, and b) establishing a mapping between the classification system (thesaurus) previously in use at Diari SEGRE, and the IPTC standard. In order to represent the actual archive contents, we have built our own ontology, which is described next.

4.3 Identification of Concepts and Class Hierarchies

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.

One of the problems that came up when the daily work of editors and documentalists was analysed is that the way in which editors search information in the database greatly differs from that in which archivists annotate and store this information. For this reason the use of the subject reference system by IPTC was proposed to archivists, who agreed that this could be a solution to the problem.

The ontology to represent the archives of Diari SEGRE has been built by using the Protégé ontology editor [21]. 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, three ways of classifying contents have been included. In one of them contents are classified by subject following the IPTC subject classification hierarchy. An alternative classification can be made according to contents genre, which has to do with the nature (breaking news, summary, interview, opinion, survey, forecast, etc.) of a news or photograph rather than its specific contents. Finally, a content can be classified according to some keywords that describe it.

These three classifications replace those by section and category in the thesaurus that were used so far by documentalists. A mapping has been established between the old thesaurus categories and those in the IPTC hierarchy. In some cases this correspondence is not one to one so contents are allowed to be classified in more than one category.

Other concepts have been created that do not correspond directly to information stored in the current database, among which it is worth mentioning the NewsRelation concept that allows establishing relationships among news such as "extension", "previous", "comment", etc. The possibility of extending the ontology with other concepts and descriptions remains open. The characteristics of the employed Semantic Web technologies are devised to facilitate this kind of extension.

5 Semantic Search and Navigation

The defined ontology serves to enrich the existing news archive contents with explicit semantic representations, giving rise to a semantic knowledge base. The ontology provides the vocabulary to express descriptions and associate them to the resources stored in the archive. The added value that results from this enrichments pays-off with the possibility to develop advanced exploitation modules like the ones developed in Neptuno, namely, a search module, and a system for ontology visualisation, both integrated in a semantic portal.

5.1 Semantic Search

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, cause an information overload due to many documents offered as the response.

A semantic search engine [22, 23] 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 modelled 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 (ontology instances).

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 (see figure 2). The interaction between both aspects of search 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 or restrict the search successively.

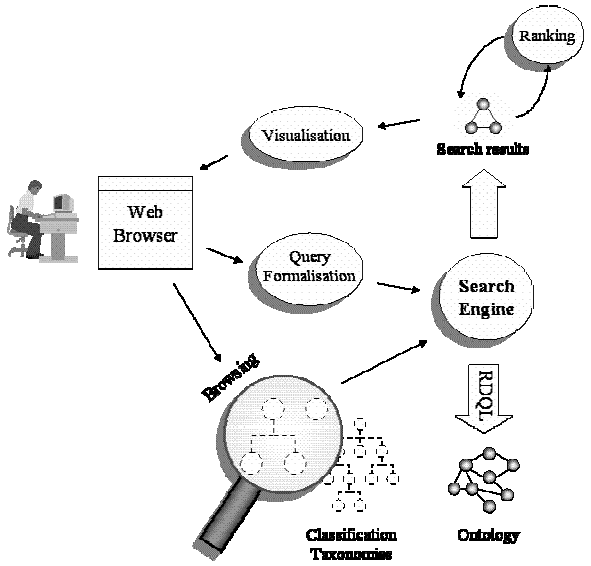

Fig. 2. Navigation and search in the Neptuno platform

5.2 Visualisation

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 visualisation criteria are used to build an ontology and often the information is not suitable to be published as it is:

- Concepts may have too many attributes.
- When relations are represented as independent concepts (first class objects) the navigation becomes tedious.
- Concepts to be shown do not always correspond to modelled ones.

In our case, in the ontology described in this paper, modelling of concepts and relations in the newspaper archive 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 visualised. That's why we introduce the concept of Visualisation Ontology [24]. This ontology, called publication schema, allows organising the concepts and attributes in order to be published in the portal.

The visualisation 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 visualised through different views.

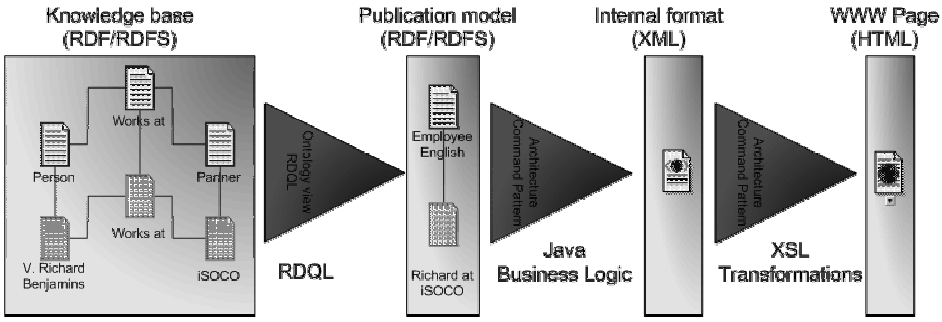

Fig. 3. Publication through visualisation ontology

The visualisation ontology has two predefined concepts:

- Publication entity: Concept that encapsulates objects as they will be published in the portal. Any concept defined in the visualisation 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 (whether the transformation style-sheet is going to translate into HTML, WAP, or just XML). In this case, the news library is exported in HTML format.

Back-office management is naturally divided into two tasks:

- Content management on domain ontology: adding new instances or modifying the overall schema.
- Visualisation management on publication ontology: modifying how information is shown (look and feel, layout, etc.)

Both tasks are performed using Protégé 2000 editor, since both domain and publication model are defined in RDF language.

6 Architecture and Implementation

In Neptuno we have undertaken the introduction of the Semantic Web proposals by following a smooth transition strategy [25], which advises to keep the compatibility (at least initially) with the current technology: browsers, protocols, web and application servers, databases, architectures.

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. At the time of this writing we are testing a first prototype. The Neptuno ontology, that is to say the hierarchy of concepts, properties and relationships, has been built manually. It contains 1,330 classes and 44 properties. 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 [26]. To limit the volume of data in our preliminary testing phase, we are initially working with a subset of the archive, including only the contents produced in 2003, which comprise 39,084 news and 19,573 photographs. From these contents a total of 95,615 instances and 857,447 sentences in RDF are automatically generated by this module.

The manual creation of instances for new informations 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 visualisation 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 formalised 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 visualisation module, which generates an HTML page where the list of found instances is displayed in an abbreviated and clickable form. When the user selects an instance, the visualisation module on the server notices 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.

Access to the knowledge base from the search and visualisation modules is carried out by means of the Jena library [27] for RDF.

7 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 actual application experience of Semantic Web technologies in a real setting, and makes 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.

Besides these immediate advantages, the work done allows undertaking now higher-level problems from the grounds established so far, which we are starting to undertake at

the time of this writing. For example, the expressiveness of the ontology developed in this phase is limited in terms of the semantics that is actually added with respect to the information already present in the current archive databases. As pointed out in Section 6, the manual introduction of this semantics is unfeasible because of its high cost, most of all for old contents. The feasible means to carry out this enrichment are those of semiautomatic kind, by means of metadata extraction modules, based on text analysis and text mining, that generate relationships between news, detect concepts in news bodies, classify contents, etc. This enrichment would enable the development of further capabilities even more sophisticated.

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 already mentioned 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 go further in the improvements achieved so far, increasing for example, the precision of the search system by means of 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.

The ease of this kind of extensions over a platform like the one developed is an important indirect advantage of the work done to this point, and an intentional feature of the used technologies.

8 Acknowledgements

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

References

1. NewsLibrary, the world's largest news archive, http://www.newslibrary.com
2. The British Library, the world's knowledge, http://www.bl.uk
3. Baumgart, J.: U. S. Newspaper Archives on the Web. Available at http://www.ibiblio.org/slanews/ internet/archives.html
4. NewsViews Solutions, http://www.newsviewsolutions.com
5. ActivePaper Archive by Olive Software, http://www.active-paper.com/ap_aparchive.html
6. LexisNexis for law, public records, company data, government, academic and business news sources, http://www.lexisnexis.com
7. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (2001).
8. IPTC Subject Reference System & NewsML Topicsets, http://www.iptc.org/metadata
9. Milenium Arcano by Protec, http://www.mileniumcrossmedia.com/Arcano/Arcano.htm

10. El País - el archivo - Hemeroteca, http://www.elpais.es/archivo/hemeroteca.html
11. ProQuest Historical Newspapers, http://www.il.proquest.com/products/pt-product-HistNews.shtml
12. ArchiveIQue by Baseview, http://www.baseview.com/products/archiveique.html
13. Canto - Digital Asset Management with Cumuluc, http://www.canto.com
14. DC4, The Digital Collections System, http://www.digitalcollections.biz/dc4.asp
15. Lassila, O., Swick, R.R.: Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. Available at http://www.w3.org/TR/REC-rdf-syntax
16. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (2001)
17. IPTC NewsML, http://www.newsml.org
18. IPTC News Industry Text Format (NITF), A Solution for Sharing News, http://www.nitf.org
19. XMLNews, XML and the News Industry, http://www.xmlnews.org
20. Publishing Requirements for Industry Standard Metadata (PRISM), http://www.prismstandard.org
21. Noy, N.F., Sintek, M., Decker, Crubezy, M., Fergerson, R.W., Musen, M.A.: Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2) (2001) 60-71
22. Guha, R., McCool, R., Miller, E.: Semantic search. 12th International World Wide Web Conference (WWW2003), Budapest, Hungary (2003), 700 – 709
23. Shah, U., Finin, T., Joshi, A., Cost, R.S., Mayfield, J:. Information Retrieval on the Semantic Web. 10th International Conference on Information and Knowledge Management (2002).
24. 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, http://www.drecommerce.com/doc/Benjamins-Duontology-a.pdf
25. Haustein, S., Pleumann, J.: Is Participation in the Semantic Web too Difficult? International Semantic Web Conference (ISWC'2002). Sardinia, Italy (2002)
26. Bizer, C.: D2R MAP - A Database to RDF Mapping Language. 12th International World Wide Web Conference (WWW2003). Budapest, Hungary (2003)
27. Jena 2 – A Semantic Web Framework, http://www.hpl.hp.com/semweb/jena2.htm