Neptuno: tecnologías de la web semántica para una hemeroteca digital

P. Castells, E. Pulido, C. Carranza, M. Rico
Universidad Autónoma de Madrid
Ctra. de Colmenar Viejo km. 15
28049 Madrid
+34 – 914 972 222
{pablo.castells,estrella.pulido, mariano.rico,cesar.carranza} @uam.es

F. Perdrix, E. Piqué, J. Cal
Diari Segre S.L.U.
C/ Del Riu nº6
25007 Lleida
+34 – 973 248 000
{fperdrix,epique,jcal} @diarisegre.com

R. Benjamins, J. Contreras
iSOCO, S.A.
c/ Fca. Delgado 11 – 2º
28100 Alcobendas – Madrid
+34 – 913 349 797
{rbenjamins,jcontreras} @isoco.com

J. Lorés, T. Granollers
Universitat de Lleida
C/ Jaume II nº69
25001 Lleida
+34 – 973 702 720
{jesus,tonig} @griho.net

RESUMEN

La hemeroteca de un periódico es una herramienta de trabajo fundamental para los equipos de redacción. Su explotación en formato digital a través de la web, y el suministro de la tecnología que lo haga posible, es también un negocio importante hoy en día. El volumen de los contenidos de una hemeroteca, y la complejidad de los equipos humanos que las crean y las mantienen, dan lugar a diversas dificultades de gestión. Proponemos la introducción de las tecnologías emergentes basadas en semántica para mejorar los procesos de creación, mantenimiento y explotación de la hemeroteca digital de un diario. Describimos una plataforma basada en estas tecnologías, que consiste en a) una base de conocimiento asociada a la hemeroteca, basada en una ontología para la descripción de información periodística, b) un módulo de búsqueda semántica, y c) un módulo de exploración y visualización de contenidos basado en ontologías.

Palabras clave

Web Semántica, Ontología, Búsqueda en la Web, Visualización de Información, Prensa Digital, Hemeroteca, Portal Semántico, Navegación Hipermedia

1. INTRODUCCIÓN

La introducción de las nuevas tecnologías de la información en el sector mediático ha marcado un nuevo ciclo evolutivo en la actividad periodística. La prensa digital permite una difusión sin precedentes, facilidad de acceso, inmediatez, economía de medios, extensión ilimitada, e información a la carta, eliminando las restricciones de tiempo y espacio, y alterando las rutinas de los equipos de redacción [9]. Se han transformado profundamente las formas de comunicación entre los distintos agentes especializados que participan en la industria mediática (empresas, medios, grupos, agencias, consorcios, profesionales, etc.), con la puesta en funcionamiento de nuevas infraestructuras, protocolos y estándares de intercambio, para la distribución y/o venta automática (push) o bajo demanda (pull) de paquetes informativos a través de distintos canales y formatos de transmisión. A nivel interno, la tendencia en los productores de información apunta a la adopción de plataformas integradas que den soporte a todo el ciclo de elaboración, gestión y publicación de contenidos, abarcando la recepción de información externa (p.e. noticias de agencia), la elaboración de contenidos propios, maquetación, documentación, gestión de archivo, etc.

Una de las consecuencias interesantes de esta transformación tecnológica en la industria mediática ha sido la emergencia, en muy pocos años, de todo un nuevo mercado de servicios online para la redistribución, sindicación, agregación, y corretaje de noticias de archivo (ver por ejemplo NewsLibrary [18], BritishLibrary [5], o una lista de hemerotecas online de periódicos estadounidenses [2]). La hemeroteca de un periódico es un producto informativo muy valioso para una amplísima variedad de consumidores: estudiantes, investigadores, historiadores, profesionales de los negocios, público en general, y no menos importante, los propios redactores de noticias. El suministro de tecnología para la construcción y mantenimiento de hemerotecas, con medios de acceso, publicación y pago, es hoy en día un negocio que mueve importantes presupuestos (véase por ejemplo, NewsViews Solutions [20], ActivePaper Archive [1]).

Las noticias recogidas de la publicación diaria en medios de comunicación representa un enorme volumen de información (p.e. LexisNexis [15] dice manejar más de 3.300 millones de documentos), muy vagamente organizada (p.e. comparada con una biblioteca de libros), que crece sin una estructura global a priori, a medida que nuevas historias se suman y evolucionan impredeciblemente. Este corpus en constante crecimiento de noticias archivadas resulta del trabajo coordinado, pero en buena medida autónomo, de un grupo de reporteros cuyo objetivo primario no es construir una hemeroteca, sino proporcionar el mejor producto informativo posible para el consumo inmediato. Los periodistas suelen estar asistidos por documentalistas que ayudan a clasificar, indexar y anotar noticias cuando éstas pasan

Interacción 2004, 3-7 mayo, 2004, Lleida (España).

al archivo, utilizando software específico para la gestión de hemerotecas.

Además de esto, se necesitan mecanismos potentes de búsqueda y exploración para que los consumidores de información puedan encontrar su camino en este espacio informativo. La tecnología actual típicamente proporciona sistemas de búsqueda basada en palabras clave (a menudo por campos: titular, resumen, cuerpo, sección), funcionalidades de exploración dentro de un ejemplar de un periódico y, en ediciones web, navegación a través de hiperenlaces creados a mano entre noticias (p.e. enlaces a antecedentes de una noticia).

Existe aún un amplio margen para aprovechar las posibilidades que ofrece el medio digital para la explotación de una hemeroteca. Los aspectos que pueden ser mejorados incluyen: a) la búsqueda basada en palabras clave resulta limitada en capacidad expresiva; b) débil interrelación entre elementos de la hemeroteca: los usuarios pueden necesitar combinar varias consultas indirectas manualmente para obtener respuestas a consultas complejas; c) falta de un estándar comúnmente adoptado para la representación y compartición de noticias entre periódicos; d) falta de consenso interno entre periodistas y documentalistas sobre la terminología de descripción de contenidos; e) falta de implicación de los redactores en el proceso de archivamiento.

Una de las novedades más prometedoras en la línea de superar el tipo de limitaciones señaladas, para un sistema de gestión de información de las características descritas, son las nuevas tecnologías emergentes de lo que se ha dado en denominar la *web semántica*, que propone nuevas técnicas, paradigmas y estándares para la representación del conocimiento que faciliten la localización, compartición e integración de recursos en la WWW [3].

El proyecto Neptuno[1] se ha puesto en marcha con el fin de aplicar estas tecnologías para mejorar el estado del arte en distintos aspectos de la producción y consumo de prensa digital. Se trata de un proyecto conjunto que agrupa a dos universidades (Universidad Autónoma de Madrid y Universitat de Lleida), un medio de comunicación (Diari SEGRE), y una empresa proveedora de tecnología (iSOCO, S. A.). En este artículo se presentan los resultados de la primera fase del proyecto, centrada en la construcción, gestión y explotación de la hemeroteca de un periódico.

El objetivo del trabajo aquí descrito es el desarrollo de una hemeroteca semántica de alta calidad para el Diari SEGRE, en la que a) los periodistas y documentalistas tienen medios más expresivos para describir y anotar el material informativo, b) se proporcionan mejores capacidades de búsqueda y exploración a los redactores y lectores que las actualmente disponibles, y c) se facilita la apertura de la hemeroteca a su integración en mercados electrónicos potenciales de productos informativos.

De acuerdo con estos objetivos, se ha desarrollado una plataforma cuyas principales componentes son:

- Una ontología para noticias de archivo, basada en la experiencia y práctica de los periodistas y documentalistas, e integrando estándares dominantes del consorcio IPTC [13].
- Una base de conocimiento donde el material de archivo se describe utilizando la ontología. Un módulo de mapeo de base de datos a la ontología integra automáticamente el material antiguo de la hemeroteca en la base de conocimiento.
- Un módulo de búsqueda semántica que permite expresar necesidades informativas en términos de la ontología, y proporcionar respuestas más precisas.
- Un módulo de visualización y navegación pra mostrar elementos individuales del archivo, partes o combinaciones de elementos, y listas o grupos de elementos

El resto de este artículo está organizado como sigue. En el próximo apartado se describe la dinámica de creación y mantenimiento de la hemeroteca en el Diari SEGRE, con la tecnología empleada actualmente. Seguidamente se presentan las propuestas genéricas de la web semántica para mejorar la gestión masiva de recursos en la web (y otros repositorios de menor escala). El apartado 4 describe la definición de una ontología para el proyecto Neptuno, y el 5 explica las funcionalidades de búsqueda y visualización sobre la base de conocimiento. El apartado 6 describe la arquitectura de la plataforma y la construcción de la base de conocimiento a partir de las bases de datos actuales, concretando algunos detalles de implementación.

2. GESTIÓN DOCUMENTAL EN EL DIARI SEGRE

La elaboración de noticias dentro de un grupo mediático como el Diari SEGRE se nutre de diversas fuentes de información de entre las que destaca el uso de las hemerotecas de noticias. Estas hemerotecas digitales constituyen siempre un referente de búsquedas de antecedentes o de otras noticias relacionadas que complementen, maticen o ayuden a situar en un cierto contexto la nueva información que el periodista está redactando.

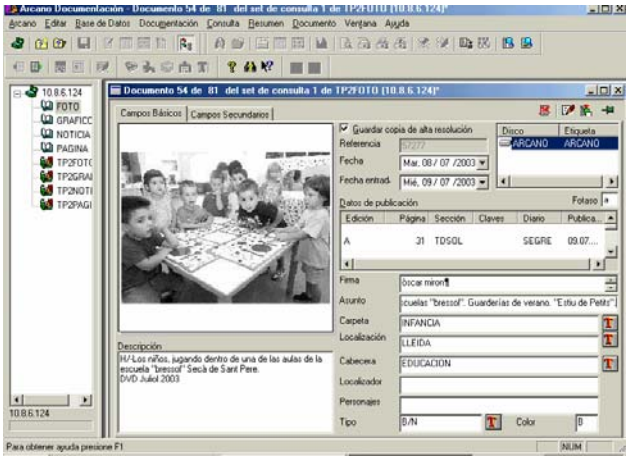

Figura 1. Entorno actual de gestión de la hemeroteca

En la mayoría de los medios periodísticos actuales existen grandes repositorios de información basados en un sistema de bases de datos relacional y un cierto software que permite

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

gestionar su contenido. En el caso del Diari SEGRE, éste tiene implantado un paquete integrado llamado Millenium Arcano [16] (ver Figura 1), uno de los más usados y populares en el sector en España y Portugal.

Este sistema precisa que cada día se archiven las noticias publicadas en la hemeroteca. El proceso de archivar las noticias se realiza desde el departamento de documentación, según los criterios de los expertos en este departamento y según las posibilidades que les permite esta plataforma software, clasificando las noticias según un tesauro jerárquico en constante evolución de los conceptos que pretende aglutinar.

Por otro lado, el consumo de esta información, almacenada previamente, lo realizan los periodistas cuando se auto-asesoran sobre ciertos temas o sucesos, no siempre con la calma que se prevé necesaria para hacerlo, ni con la suficiente información de antemano como para cerciorarse de que buscan algo en concreto.

Es por todo ello que la diferencia de los modelos mentales (puesta en evidencia en el transcurso de este proyecto) entre los/las documentalistas y los/las periodistas hacen necesario articular un sistema más flexible de categorización y de búsqueda de contenidos, que es lo que se pretende lograr en el proyecto Neptuno, ayudándonos de las posibilidades que nos ofrece la web semántica.

3. LA WEB SEMÁNTICA

La web semántica es un área pujante en la confluencia de la Inteligencia Artificial y las tecnologías web, que propone la introducción de descripciones semánticas explícitas de los contenidos en la web para facilitar su localización, compartición e integración [3]. Se trata de describir los recursos con representaciones procesables no sólo por personas, sino por máquinas que puedan asistir, representar, o reemplazar a las personas en tareas rutinarias o inabarcables para un humano.

La web semántica mantiene los principios que han hecho un éxito de la web actual, como son los principios de descentralización, compartición, compatibilidad, o la apertura al crecimiento y uso no previstos de antemano. En este contexto un problema clave es alcanzar un entendimiento entre las partes: usuarios, desarrolladores y programas de muy diverso perfil. La web semántica rescata la noción de ontología del campo de la Inteligencia Artificial como vehículo para cumplir este objetivo [10].

Una ontología es una jerarquía de conceptos con atributos y relaciones, que proporciona un vocabulario consensuado para definir redes semánticas de unidades de información interrelacionadas. Durante los últimos años se han desarrollado diversos lenguajes y estándares para la definición de ontologías, entre ellos XML, RDF [14], DAML+OIL [6], y más recientemente OWL [8], respaldados por el consorcio W3C, uno de los principales promotores de la web semántica.

Existe un gran interés desde el entorno corporativo, el sector público y el mundo académico por hacer de la web semántica una realidad, ya que se piensa que puede ser una pieza importante para el progreso de la sociedad de la información. Para ello se esta invirtiendo un gran esfuerzo en desarrollar a) la infraestructura necesaria para su despliegue, b) aplicaciones que demuestren la viabilidad y el beneficio de la web semántica y a la vez motiven el desarrollo y consumo de infraestructura y c) nuevas soluciones para resolver problemas específicos, e ideas que mejoren, amplíen y/o exploten las posibilidades de la web semántica.

Si bien la mayoría de grupos y empresas que participan en este campo concuerdan en el gran potencial de estas tecnologías, los resultados alcanzados hasta ahora son muy preliminares si se mira desde la óptica más ambiciosa, la de la adopción universal de la web semántica. El desarrollo de infraestructura y tecnología para la construcción de la web semántica está hoy en un punto bastante avanzado, sin embargo el uso que se está haciendo de ella no ha alcanzado aún un grado de desarrollo comparable. El desarrollo de aplicaciones que hagan uso de esta tecnología ha sido identificado como una de las realizaciones prioritarias en este punto para el progreso de la web semántica [12]. El trabajo que aquí se presenta representa una contribución que responde esta demanda, sirviendo de experiencia piloto en un dominio concreto, el periodístico.

4. ONTOLOGÍA PARA UNA HEMEROTECA

El primer paso en el desarrollo del proyecto ha sido la definición de una ontología para representar y procesar información periodística. Tras evaluar los lenguajes y estándares disponibles para la definición de ontologías, hemos optado por RDF [14], actualmente el estándar más maduro, consolidado y extendido en los últimos desarrollos y proyectos del área de la web semántica.

De acuerdo con las metodologías de referencia en la web semántica, los pasos recomendados para la construcción de una ontología son [21]:

1. Determinar el dominio y ámbito de la ontología.
2. Determinar la intención de uso de la ontología.
3. Reutilizar ontologías o vocabularios controlados existentes.
4. Enumerar los términos importantes del dominio.
5. Definir la jerarquía de clases.
6. Crear las instancias.

En Neptuno nos hemos guiado por esta recomendación, adaptándola a nuestro caso particular sin desviaciones significativas. A continuación describimos el proceso de definición en cada uno de estos aspectos, y la ontología resultante. El último de estos pasos, la creación de instancias, que en Neptuno se realiza automáticamente, se describe más adelante, en el apartado 6 sobre la arquitectura de la plataforma.

4.1 Dominio, ámbito e intención de uso

Aunque aparentemente la selección del dominio sería una cuestión obvia, el campo del periodismo tiene la peculiaridad de abordar potencialmente temas de todos los ámbitos de la actualidad y el conocimiento humano: política, cultura, tribunales, ciencia, deporte, arte, economía, etc. Ha sido necesario por ello establecer cuidadosamente un límite de los dominios a representar, sin el que cualquier intento de completitud daría lugar a un proyecto entero para cada área temática informativa.

El reflejo conceptual de las áreas temáticas en nuestra ontología se limita a la definición de categorías genéricas por temas y

subtemas, como “política”, “inmigración”, “economía”, “comercio”, “bolsa”, o “deportes”, como se describirá a continuación, pero no se incluyen clases y entidades específicas para estas áreas, como podrían ser “partido político”, “sufragio”, “juez”, “abogado”, “sentencia”, “deportista”, “actor”, “obra de teatro”, o “grupo musical”, ni las instancias de estos tipos de entidad. La creación de una base de conocimiento para estas entidades, que cubra completamente el ámbito informativo potencial de un medio de comunicación, excede la capacidad de cualquier organización, empresa o agencia que se propusiera semejante tarea. Por otra parte la utilidad de una colección parcial, de construcción más viable, es difícil de justificar.

En cuanto a la intención de uso, tras un análisis detenido, hemos llegado a la conclusión que la gestión de la hemeroteca es, de todos los ámbitos potencialmente abordables en el ciclo productivo de un periódico, posiblemente el que más y mejores oportunidades ofrece de mejorar los procesos y productos, y más se presta a las propuestas de la web semántica. Entre otros motivos que fundamentan esta valoración, podemos citar a) la calidad de los contenidos de la hemeroteca, que han pasado un filtro de selección (p.e. frente al flujo de noticias que diariamente llega a la redacción), dado que únicamente las noticias que se incluyen en las páginas del diario pasan a la hemeroteca; b) el enriquecimiento de los contenidos con metadatos, descripciones y una catalogación manual cuidadosa por parte de los documentalistas que supervisan el tránsito de las noticias a la heroteca una por una; y c) la permanencia del material por un período de tiempo indefinido (una vez más, frente al material de agencia y otras fuentes que finalmente se descartan), lo que permite una explotación continuada del valor añadido resultante de las aportaciones que aquí se proponen.

4.2 Ontologías existentes y vocabularios controlados

Nuestra primera constatación, a la hora de considerar la reutilización de ontologías y estándares en este campo, es que no se ha publicado hasta la fecha una ontología de periodismo propiamente dicha. En este sentido nuestro trabajo representa una contribución al crecimiento de la web semántica y las colecciones de ontologías de dominio público.

Por lo que respecta a otro tipo de vocabularios controlados, en el campo del periodismo se han desarrollado diferentes estándares, entre los que cabe mencionar NewsML [19], NITF [17], XMLNews [25], la jerarquía de clasificación por temas del IPTC [13], y PRISM [23]. Tanto NewsML como NITF (News Industry Text Format) son estándares basados en XML para representar y gestionar noticias durante todo su ciclo de vida, incluyendo la creación, intercambio y consumo de las mismas, pero mientras que NewsML se utiliza para representar noticias como paquetes multimedia, NITF se ocupa de la estructura de los documentos. XMLNews es un subconjunto de NITF y está basado en RDF. Incluye una serie de marcadores (como localización, personaje o fecha) que permiten anotar noticias para facilitar la búsqueda de información.

Estos tres estándares han sido creados por IPTC (International Press Telecommunications Council), un consorcio internacional de agencias de noticias, editores y distribuidores de periódicos. Este organismo ha propuesto el Subject Reference System [13], una jerarquía de clasificación por temas en tres niveles con diecisiete categorías para el primer nivel.

PRISM (Publishing Requirements for Industry Standard Metadata) fue desarrollado por IDEAlliance (International Digital Enterprise Alliance), un consorcio industrial de empresas editoriales y de fabricantes de software de publicación entre las que se incluyen Adobe, Quark, Condé Nast y Time Inc. Es un estándar XML parecido a NewsML de un vocabulario de metadatos para la industria editorial que facilita la agregación y sindicación de contenido digital.

Tras evaluar todos estos estándares, en Neptuno se ha adoptado el IPTC Subject Reference System como sistema de clasificación temática de contenidos. El trabajo de integración y adaptación de este estándar a nuestra ontología ha consistido en a) convertir la jerarquía temática del IPTC a una jerarquía de clases en RDF, y b) establecer un mapeo entre el sistema de clasificación (tesauro) actual en uso en la hemeroteca del Diari SEGRE y el estándar IPTC. Para la representación de los propios contenidos, se ha elaborado una ontología propia que se detalla a continuación.

4.3 Identificación de términos y jerarquías de clases

El Diari SEGRE dispone de una base de datos donde se almacena información sobre noticias, fotografías, gráficos y páginas del diario. Los documentalistas clasifican diariamente los contenidos del periódico utilizando la herramienta Milleniu, Arcano. Los criterios que se utilizan para esta clasificación son fundamentalmente dos: la sección a la que pertenece el contenido (Deportes, Sociedad, etc.) y el tema del que trata. Para esta última clasificación se utilizan un tesauro que se ha ido elaborando incrementalmente entre todos los documentalistas a lo largo del tiempo, actualizándolo conforme surgen nuevas necesidades.

Uno de los problemas planteados al analizar el trabajo diario de redactores y documentalistas es que el modo en que los redactores buscan la información en la base de datos difiere mucho de la forma en que los documentalistas archivan dicha información. Por este motivo se propuso a los documentalistas utilizar una clasificación basada en la jerarquía de temas del IPTC y éstos estuvieron de acuerdo en que podría ser una solución al problema.

La ontología para representar los archivos documentales del Diari SEGRE se ha construido utilizando el editor de ontologías Protégé [22]. Algunos de los conceptos de esta ontología corresponden a las tablas de la base de datos existentes, como por ejemplo, Noticia, Foto, Gráfico y Página. Todos estos conceptos son subclases del concepto Contenido.

Además se han incluido tres formas de clasificar un contenido. En una de ellas se clasifican los contenidos de acuerdo al tema siguiendo la jerarquía de clasificación por temas del IPTC. Una clasificación alternativa puede realizarse de acuerdo al género que describe la naturaleza o característica de una noticia o fotografía y no su contenido específico. Por último, un contenido puede clasificarse de acuerdo a palabras clave que lo describen. Estas tres clasificaciones sustituyen a las clasificaciones por sección y por categoría de tesauros que utilizaban hasta ahora los documentalistas. Se ha establecido una correspondencia entre estas categorías y las que aparecen en la clasificación por temas

del IPTC. En algunos casos esta correspondencia no es biunívoca por lo que se permite que un contenido esté clasificado por mas de un tema.

Se han creado también otros conceptos que no se corresponden directamente con información almacenada en la base de datos actual y entre los que cabe destacar el de RelacionNoticia con el que se establecerían relaciones entre noticias como las de ampliación, antecedente, comentario, etc. Queda abierta la posibilidad de extender la ontología con otros conceptos y descripciones. Las características de las tecnologías de la web semántica utilizadas están ideadas para facilitar este tipo de extensión.

5. NAVEGACIÓN Y BÚSQUEDA

La ontología definida sirve para enriquecer los contenidos de la hemeroteca con representaciones semánticas explícitas, dando lugar a una base de conocimiento. La ontología proporciona el vocabulario para expresar las descripciones y asociarlas a los recursos almacenados en la hemeroteca. El valor añadido que resulta de este enriquecimiento se materializa en la posibilidad de desarrollar módulos de explotación avanzada como los que se han construido en Neptuno, concretamente, un módulo de búsqueda y un sistema de visualización para ontologías, integrados en un portal semántico.

5.1 Búsqueda semántica

Por lo que respecta a la búsqueda, la disponibilidad de información semántica en la base de conocimiento permite formular al usuario consultas más expresivas y precisas, e implementar un sistema capaz de utilizar elementos conceptuales para determinar correspondencias entre consultas y contenidos. La mayoría de los buscadores que actualmente existen están basados en palabras clave. El usuario puede introducir las palabras relevantes y el buscador recupera todos los documentos que las contienen: se buscan ocurrencias de palabras en documentos sin tener en cuenta:

- El significado de las palabras (pueden tener múltiples).
- Las relaciones de las palabras que forman la consulta .

Lo cual puede resultar en los siguientes problemas:

- La presentación de muchos documentos con poca relevancia con respecto a la consulta original.
- Es responsabilidad de la persona abrir cada documento para comprobar su relevancia.

La consecuencia de estos problemas puede ser que los usuarios no encuentren la información que buscan, aunque exista, o al contrario, encuentran demasiada información.

Un buscador semántico [11, 24] tiene conocimiento del dominio en cuestión. La presencia de una ontología del dominio que estructura y relaciona la información de acuerdo a su significado permite construir un buscador donde los usuario especifican sus criterios de búsquedas en función de los conceptos y atributos modelizados. Los resultados se presentan en una forma estructurada incluyendo solamente la información solicitadas. A diferencia de buscadores tradicionales donde la respuesta consiste en documentos enteros, para que sea el propio usuario que encuentre la respuesta manualmente, los buscadores semánticos devuelven solamente la información solicitada.

El módulo de búsqueda de Neptuno se ha desarrollado siguiendo estos principios de búsqueda semántica. Además, el buscador combina las consultas directas por clases de contenidos y campos de las clases, con la posibilidad de explorar de la taxonomía IPTC de clasificación temática de las noticias y documentos de la hemeroteca (ver Figura 2). La interacción entre ambos aspectos de la búsqueda es doble. Por un lado, la búsqueda por campos se restringe a las categorías del IPTC seleccionadas por el usuario. Por otro, Neptuno muestra la lista de categorías a las que pertenecen los resultados de una consulta, para que el usuario pueda navegar directamente a ellas, o seleccionarlas para restringir o ampliar sucesivamente la búsqueda.

Figura 2. Búsqueda y visualización en Neptuno

5.2 Visualización

A pesar de las ventajas que los modelos semánticos presentan a la hora de recuperar información, uno de los problemas de estos modelos es su presentación legible y comprensible al usuario final. En la fase de diseño y construcción de un modelo se valora su poder expresivo y no se toman en cuenta aspectos visuales o estéticos del mismo. Este también ha sido el caso de la ontología que se presenta en este trabajo. La modelización de los conceptos y relaciones de la hemeroteca del diario no se ha visto restringida por criterios de publicación de su contenido como podrían ser el numero de atributos de un concepto, numero de instancias o existencia de conceptos auxiliares para la representación de relaciones.

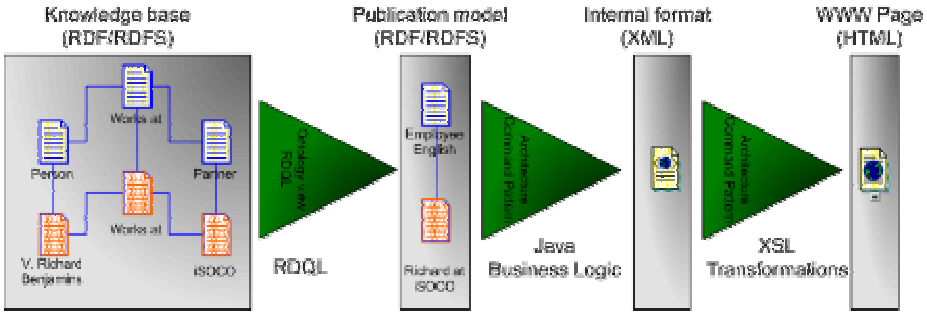

Figura 3. Fases de publicación

El modelo de publicación utilizado para la ontología de la hemeroteca se basa en la existencia de una ontología auxiliar (ontología de visualización) que permite definir vistas sobre el modelo de la hemeroteca (ontología de dominio) [7]. Estas vistas se definen de acuerdo con criterios de usabilidad y estéticos con el fin de garantizar una presentación legible del modelo semántico subyacente.

La ontología de visualización es un contenedor de entidades y atributos publicables que extraen los valores mediante lenguajes de consultas sobre RDF. Estas entidades publicables contienen aquellos atributos de la ontología de dominio que se van a presentar al usuario final. El contenido de una instancia publicable puede agrupar varios conceptos del modelo semántico original, o al contrario, puede dividir un concepto complejo en varias entidades visualizables por separado.

Más concretamente, la ontología de visualización incluye dos conceptos predefinidos:

- *Entidad de publicación*: concepto que encapsula objetos tal como se verán publicados. Todo concepto definido en la ontología de publicación heredará de él y deberá definir los siguientes atributos:
  - *Hoja de estilo* XSL asociada al concepto, que traduce sus instancias a HTML.
  - *Consulta* RDQL que obtiene todos los valores de los atributos de la instancia correspondiente en la ontología del dominio.
- *Slot de publicación*: todos los atributos que se hayan de mostrar en la web deben heredar de este concepto. La forma en que el atributo se muestre en la página web se define mediante las siguientes propiedades:
  - *Etiqueta web*: la etiqueta que aparecerá con el valor del slot.
  - *Rdql*: consulta a ejecutar para obtener el valor del slot.
  - *Link*: si el valor publicado debe realizar alguna acción al pulsar sobre él (hiperenlace, mail, botón, etc.), la acción se describe aquí.

Las componentes de una página web para visualizar instancias de la ontología se describen como subclases de *Entidad de publicación*, y sus instancias se definen de acuerdo con el canal de publicación (HTML, WAP, VoiceML, XML) a generar, a través de transformaciones XSL (ver Figura 3). En el caso de la hemeroteca, el canal de destino es la web, con lo cual el resultado de las transformaciones son páginas HMTL.

La administración de portales donde se ha empleado el uso de esta tecnología se divide así de manera natural en dos partes:

- Administración del contenido: gestiones sobre la ontología de dominio.
- Administración de la presentación: gestiones sobre la ontología de visualización que permiten modificar la agrupación de la información presentada así como su estética.

La separación entre la ontología para la representación del conocimiento del dominio y la ontología de visualización facilita la independencia de estas dos labores. Otra ventaja de la aproximación seguida es que para ambas se pueden utilizar las mismas herramientas de manejo de lenguajes de ontologías, como Protégé.

6. ARQUITECTURA E IMPLEMENTACIÓN

En Neptuno hemos abordado la introducción de las propuestas de la web semántica sigiendo la estrategia de una transición suave [12], que aconseja mantener la compatibilidad (al menos inicialmente) con la tecnología actual: navegadores, protocolos, servidores web y de aplicaciones, bases de datos, arquitecturas.

La versión actual de la plataforma Neptuno es una extensión de los sistemas previamente en funcionamiento en el Diari SEGRE que no interfiere con la tecnología y usos actualmente en funcionamiento. La ontología Neptuno, es decir la jerarquía de conceptos, propiedades y relaciones, ha sido construida manualmente. La creación manual de instancias de la ontología, asociadas a varios millones de noticias acumuladas durante los últimos años años en la hemeroteca del Diari SEGRE, sería sin duda la mejor manera de asegurar la calidad de la semántica introducida, pero también un trabajo desproporcionado e inviable, fuera de toda relación coste / beneficio razonable.

Por este motivo Neptuno incluye un módulo que puebla la ontología con instancias que se extraen automáticamente de las bases de datos del sistema Arcano, utilizando una herramienta de mapeo de bases de datos JDBC/ODBC a RDF [4]. La creación manual de instancias para las nuevas noticias y fotografías que se almacenen en la hemeroteca, a un ritmo diario normal, a patir de la implantación de Neptuno sí es una meta abordable en un futuro. La introducción de nuevas herramientas de documentación

semántica requiere sin embargo un trabajo cuidadoso de análisis, diseño, testeo y contrapeso del esfuerzo adicional que puedan acarrear tales herramientas al trabajo de los documentalistas. Entretanto hemos decidido no interferir en los entornos actuales con los que interactúan los trabajadores del Diari, y generar diariamente las instancias por el mismo procedimiento automático utilizado con el material antiguo, sin alterar el ritmo de producción del periódico.

Los módulos de búsqueda y visualización de contenidos operan directamente sobre la base de conocimiento en RDF. El usuario formula búsquedas mediante una interfaz web en la que selecciona la clase de contenidos a buscar (Noticia, Fotografía, Gráfico o Página), y especifica palabras clave para los campos que desee (titular, autor, sección, fecha, tema, etc.) de la clase seleccionada. Esta información es enviada al servidor de Neptuno, donde se formaliza la petición bajo la forma de una consulta RDQL. Esta consulta se ejecuta sobre la base de conocimiento, obteniéndose como resultado una lista de recursos (instancias) que verifican las condiciones de la consulta.

La lista de recursos es enviada al módulo de visualización, que devuelve una página HTML donde se muestra al usuario la lista de instancias encontradas, en forma abreviada y seleccionable. Cuando el usuario selecciona una instancia, el módulo de visualización en el servidor observa la clase a la que pertenece el recurso, selecciona la entidad de publicación correspondiente, y genera una página web donde se muestran los detalles del recurso, según los principios descritos en el apartado anterior.

El acceso a la base de conocimiento desde los módulos de búsqueda y visualización se lleva a cabo mediante la librería Jena para RDF.

7. CONCLUSIONES Y TRABAJO FUTURO

La hemeroteca de un periódico es una herramienta de trabajo fundamental para los equipos de redacción, y un producto potencialmente comercializable hacia diversos tipos de consumidores por distintos canales de distribución.

El tamaño y complejidad de la información almacenada, y las limitaciones de tiempo a la hora de catalogar, describir y ordenar la información entrante, hacen por lo general de las hemerotecas un corpus relativamente desorganizado y difícil de manejar. Esto las aproxima en ese sentido a las características de la web en su conjunto, por lo que comparten muchos de los problemas de ésta, y son por tanto pertinentes aquí las soluciones ideadas para la web, en particular las que propone la web semántica.

El trabajo desarrollado hasta aquí representa una experiencia de aplicación de las tecnologías de la web semántica en un escenario real, y aporta contribuciones novedosas en varios de los aspectos abordados: definición de ontologías en un dominio concreto, funcionalidades de búsqueda semántica y exploración, desarrollo de una interfaz de usuario para interactuar con la base de conocimiento, transición desde un sistema en funcionamiento con tecnologías tradicionales hacia una plataforma basada en semántica.

Además de estas ventajas inmediatas, el trabajo realizado permite abordar ahora otros problemas de mayor nivel sobre las bases sentadas hasta aquí, que nos planteamos emprender en el futuro inmediato. Por ejemplo, la expresividad de la ontología desarrollada en esta fase es limitada en cuanto a la semántica que efectivamente se añade respecto de la información ya presente en las bases de datos actuales. Como se ha señalado en el apartado 6, la introducción manual de esta semántica es inviable por excesivamente costosa, sobre todo por lo que se refiere a los contenidos antiguos. Las formas viables de realizar este enriquecimiento son las de tipo semi-automático, llevadas a cabo por módulos de extracción de metadatos mediante análisis y minería de texto, que generen relaciones entre noticias, detecten conceptos en el cuerpo de las noticias, clasifiquen contenidos, etc. Este enriquecimiento daría pie al desarrollo de nuevas prestaciones aún más sofisticadas.

Otros objetivos que estamos contemplando a partir de aquí son la integración de nuevas fuentes (hemerotecas externas), géneros de contenido (texto, audio, vídeo) e idiomas (castellano y catalán), el enriquecimiento ya citado de las descripciones de contenidos por métodos automáticos de análisis de texto y multimedia, y la adaptación automática a múltiples dispositivos y canales de acceso. Prevemos también profundizar en las mejoras conseguidas hasta aquí, aumentando por ejemplo, la precisión del sistema de búsqueda mediante algoritmos de ranking, aumentando la expresividad del mismo con interfaces más avanzadas, y llevando cabo una revisión metodológica de la representación y sistemas de clasificación (ontologías) utilizados actualmente en la base de conocimiento.

La facilidad para este tipo de extensiones sobre una plataforma como la que hemos desarrollado es una ventaja indirecta importante del trabajo realizado hasta este punto, y una característica expresa de las tecnologías empleadas.

8. AGRADECIMIENTOS

El trabajo presentado en este artículo está financiado por el Ministerio de Ciencia y Tecnología, proyectos nº FIT-150500-2003-511 y TIC2002-1948

9. REFERENCIAS

[1] ActivePaper Archive, Olive Software, http://www.active-paper.com/ap_aparchive.html.

[2] J. Baumgart, U.S. Newspaper Archives on the Web, http://www.ibiblio.org/slanews/internet/archives.html.

[3] T. Berners-Lee, J. Hendler, O Lassila. The Semantic Web. Scientific American, May 2001.

[4] C. Bizer. D2R MAP - A Database to RDF Mapping Language. 12th International World Wide Web Conference (WWW2003). Budapest, Hungary, May 2003.

[5] The British Library, http://www.uk.olivesoftware.com/.

[6] D. Connolly, F. van Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. DAML+OIL Reference Description. W3C Note 18 December 2001. Available at http://www.w3.org/TR/daml+oil-reference.

[7] J. Contreras, V. R. Benjamins, J. A. Prieto, D. Patón, S. Losada, and D. González. Duontology: an Approach to Semantic Portals based on a Domain and Visualization Ontology. KTWeb, http://www.drecommerce.com/doc/Benjamins-Duontology-a.pdf.

[8] M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language 1.0 Reference W3C Working Draft 29 July 2002. Available at http://www.w3.org/TR/owl-ref.

[9] Q. Gil. *Diseñando el periodista digital*. Revista electrónica Sala de Prensa, nº 13, noviembre 1999. Disponible en http://www.saladeprensa.org.

[10] T. R. Gruber. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), pp. 199-220, 1993.

[11] R. Guha, R. McCool, E. Miller. Semantic search. 12th International World Wide Web Conference (WWW2003), Budapest, Hungary, 2003, pp. 700 – 709.

[12] S. Haustein and J. Pleumann. Is Participation in the Semantic Web too Difficult? International Semantic Web Conference (ISWC'2002). Sardinia, Italy, 2002.

[13] IPTC Subject Reference System & NewsML Topicsets, http://www.iptc.org/metadata.

[14] O. Lassila, R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. Available at http://www.w3.org/TR/REC-rdf-syntax.

[15] The LexisNexis Group, http://www.lexisnexis.com.

[16] Millenium Arcano. http://www.mileniumcrossmedia.com/Arcano/Arcano.htm.

[17] News Industry Text Format (NITF), http://www.nitf.org.

[18] NewsLibrary, http://www.newslibrary.com.

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

[20] NewsViews Solutions, http://www.newsviewsolutions.com.

[21] N. F. Noy and D. L. McGuinness. `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, March 2001.

[22] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, & M. A. Musen. Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2), pp. 60-71, 2001.

[23] Publishing Requirements for Industry Standard Metadata (PRISM), http://www.prismstandard.org/.

[24] U. Shah, T. Finin, A. Joshi, R. S. Cost and J. Mayfield. Information Retrieval on the Semantic Web. 10th International Conference on Information and Knowledge Management, November 2002.

[25] XMLNews, http://www.xmlnews.org.