La Hemeroteca digital de El País

Flora Sanz Calama

Jefa de Documentación de Prisacom
e-mail: flora.sanz@prisacom.com

Resumen. La hemeroteca digital de El País comprende desde mayo de 1976 hasta la fecha. El presente trabajo expone el proceso de migración de los contenidos desde la base de datos de El País (Hércules) al sistema editorial de Prisacom y su conversión a XML. Se aborda también la fase de traducción del thesaurus propio de El País al usado en el sistema editorial de Prisacom, basado en la clasificación de materias del International Press Telecommunication Council (IPTC) y con incorporación de descriptores habituales en el contexto español. Por último, se expone la integración del material multimedia (imágenes de portadas y fotos) y la composición y edición de dichas portadas para su publicación en la web.

Palabras claves: Hemerotecas digitales, Prensa digital, XML, IPTC, Conversión de formatos

1. Introducción

La Hemeroteca Digital de El País comprende desde el 4 de mayo de 1976, fecha en la que comienza a publicarse el diario, hasta nuestros días. En este trabajo se aborda el proyecto de migración de los contenidos de la base de datos de El País, denominada Hércules, al sistema editorial de Prisacom para constituir una hemeroteca digital y los procesos de adaptación a nivel de clasificación y de formato realizados. Primeramente se expone la definición y estructura de ambos sistemas para clarificar el proceso de extracción y conversión y, posteriormente se presentan las fases seguidas en el proyecto.

2. La Base de Datos de El País

La base de datos de El País, denominada Hércules, es una base documental diseñada especialmente para el periódico que incorpora todo el fondo producido por el diario (texto, fotos, gráficos y páginas) clasificado en estos cuatro apartados. Actualmente, se ha desarrollado una versión, denominada Pegaso, para acceder a través de la Intranet a consultar los contenidos.

La estructura de registro de la base de datos es muy completa y con campos comunes independientemente del material, para poder ejecutar búsquedas conjuntas entre ciertos contenidos como texto y fotografías.

La definición de los registros, excepto los de página, se compone de un apartado de análisis formal y otro de contenido y los datos que se registran son los siguientes:

1. Descripción formal

- Datos de publicación: fecha de publicación, edición (primera, segunda, Nacional...), cuaderno (regionales , suplementos y extras) y sección (las que tiene el diario en su edición impresa).

- Datos de producción: fecha en la que se ha creado el documento que no tiene porque coincidir con la de publicación, especialmente en los materiales especiales (fotografía e infografía).

- Estructura del objeto: Título, Antetítulo, Firma y Lugar , Entradilla, Cuerpo de la noticia o documento gráfico. Cuenta, además, con una apartado para el Copyright, es decir, para conocer que derechos de reproducción o de comercialización se tienen sobre ese documento.

- Tipología Documental: Es la identificación del objeto textual según su género periodístico, información esencial para conseguir resultados precisos a la hora de ejecutar una búsqueda. Se distinguen los siguientes: Editorial, Opinión, Crítica, Cartas, Crónica, Fe de Errores, Entrevista, Perfil, Reportaje, Breve, Noticia, Apoyo Documental, Necrológica y Tipología indefinida.

En el caso de las fotografías se utilizan otro tipo de calificadores como retrato, plano medio, primer plano, un primer plano, etc. También, se registra si es color o blanco y negro y si es horizontal o vertical.

Las páginas, están almacenadas en formato Tiff con una "carcasa" en PDF hasta el año 2001 y desde ese año en PDF. La descripción formal se limita a asignar fecha de publicación, edición, cuaderno y sección

2. Descripción de contenido

Para realizar el tratamiento documental atendiendo a su temática, el sistema dispone de un thesaurus propio con aproximadamente 750.000 entradas organizadas en tres niveles por grandes categorías temáticas (temas) y subgrupos de temas que a su vez se relacionan con grupos de subtemas y sus subgrupos. Las entradas principales son:

- CULTURA
- DEPORTES
- ECONOMÍA Y TRABAJO

- EDUCACIÓN
- EMPRESAS
- ESPAÑA
- INTERNACIONAL
- PERSONAS
- SOCIEDAD
- SUCESOS

Por ejemplo, el tema PERSONAS tiene a su vez subtemas organizados por categorías profesionales y el tema CULTURA dispone de subtemas: CINE, ARTE, TEATRO... que a su vez, se relacionan con subgrupos de temas como PREMIOS DE CINE: Oscar, Goya, César o FESTIVALES como Festival de Cine de San Sebastián, Festival de Cannes, etc.

3. El Sistema Editorial y Documental de Prisacom

La herramienta editorial creada en Prisacom es global en el sentido de que tanto la producción y edición de los contenidos como el tratamiento documental se realiza en la misma herramienta.

El sistema está desarrollado en XML (Extensible Markup Language) [1]. Este lenguaje permite describir mediante marcas los aspectos formales y el contenido de un documento identificando los diferentes "campos informativos" que puede tener una noticia o fotografía, así se distingue una etiqueta Título, Firma, Fecha de Publicación, Sección, etc.

El estándar usado para definir la estructura del documento o DTD (Definición de Tipo de Documento) es NIFT (News Industry Text Format) [2], desarrollado por el International Press Telecommunications Council (IPTC) [3]. Las ventajas de optar por este estándar consensuado y abierto son evidentes como la posibilidad de intercambiar información en diferentes lenguas con otros medios e incorporar registros de bases de datos como es caso que se va abordar de la hemeroteca.

1 Elementos o Etiquetas Formales de NITF

La DTD de NIFT contempla sólo los elementos que componen una noticia pero adecuando ciertas etiquetas y añadiendo algunas otras específicas, se pueden definir la estructura de los objetos multimedia. De esta forma, se muestran las etiquetas utilizadas en el sistema editorial distinguiéndose los contenidos textuales y los multimedia.

1. 1 Contenidos Textuales

Se diferencian dos tipos de contenidos textuales, artículo o noticia y ficha informativa de un partido de fútbol, libro, concierto, etc.

Una noticia contempla los siguientes elementos:

- Área de Publicación

- Fecha de Publicación del documento: Año-Mes-Día
- Publicación:
- Edición: Madrid, Nacional, Barcelona, País Vasco, Andalucía, Valencia, Herald Tribune.
- Cuaderno: Regionales, Suplementos y Extras
- Sección: las del diario y suplementos

- Área del Titular:

- Epígrafe de Página: Se define como aquella información, que encabeza la página y que identifica las noticias que se publican en la misma. Generalmente es una creación periodística que engloba noticias de una misma temática dentro de una sección, suelen ser temas de permanente actualidad, por ejemplo "Crisis en Oriente Próximo" y "Los problemas de los inmigrantes".

- Epígrafe de Noticia: Actúa como el anterior pero afecta sólo a una noticia. Ejemplos claros serían en una entrevista Nombre del entrevistado y función. Además, siguiendo la propuesta de NIFT para identificar elementos, los epígrafes de página y noticia los podemos clasificar atendiendo a su contenido:

  General (Cintillo)
  Fijo (Revista de Prensa, Visto y Oído),
  Persona (Nombre)
  Función (Cargo que ostenta)
  Organización (Institución, Empresa...)
  Materia (Baloncesto, Danza...)
  Evento (Festivales, competiciones deportivas...)
  Objeto (marcas y modelos)
  Lugar

- Título
- Antetítulo
- Subtítulo
- Sumario: Utilizado en suplementos para publicar ciertos datos, que no forman parte del artículo, en la portada.
- Destacado: Frase, resaltada con otra tipografía, entresacada del texto y que se publica en un cuadro en el cuerpo de la noticia.

- Área de Firma

- Autor: nombre y apellidos
- Función: por ejemplo corresponsal, enviado especial

- Medio
- Lugar: Ciudad donde se escribe el artículo.
- Correo electrónico

• Cuerpo del documento

- Presentación: Aparece encima del titular en aquellas páginas cuyos textos se vinculen a un solo tema. Este párrafo no sustituye a la entrada o lead.
- Entradilla Larga: Texto o entrada, en negrita, que antecede al cuerpo de la noticia.
- Entradilla Corta: redactada para publicar en la portadilla de sección a partir de la anterior .
- Texto: Cuerpo de la noticia
- Pie de página:
- Despiece: Piezas sin firma que acompañan y completan un artículo.

La estructura de la Ficha es muy simple ya que sólo contempla los siguientes elementos: Epígrafe, Título y Texto.

1.2 Contenidos multimedia

En este apartado distinguimos fotografía, gráficos (estáticos y animados), PDF, audio y vídeo.

• Área de publicación:

Al ser un objeto asociado a una noticia, automáticamente incorpora los datos de publicación de la noticia.

• Área del Titular:

- Título
- Antetítulo

• Área de Firma

- Autor
- Función
- Medio
- Lugar

• Cuerpo:
- Descripción: Elemento dónde se expone brevemente el contenido del documento, curiosidades, etc.

• Características físicas de cada tipo multimedia:

FOTO:

- Color (Color ; Blanco/Negro)
- Tipo (Grande 310 x222; Pequeña 150x107)
- Orientación (Horizontal ; Vertical)
- Tamaño Bytes
- Tamaño Pixels

GRÁFICO:

1.Estático :
- Tipo (Grande 310 x222; Pequeña 150x107)
- Tamaño Bytes
- Tamaño Pixels

2. Animado

PDF:

- Archivo PDF:
- Tamaño en bytes:

AUDIO:

- Duración: Expresada en minutos y segundos.
- Codificación:
- Formato: AVI, RA
- Tamaño en bytes:

VIDEO: (Color ; Blanco/Negro)

- Duración: Expresada en minutos y segundos.
- Codificación:
- Formato: WM
- Tamaño en bytes:

2. Descripción de contenido: Clasificación

Los elementos de documentación son comunes para todos los tipos de materiales y comprende tres apartados:

- Género: Noticia, noticia de apoyo, editorial, tribuna, columnista, cartas, perfil / biografía, necrológica, entrevista, reportaje, crítica, crónica, breve, fragmento literario, fotonoticia, análisis y sin definir.

- Clasificación temática basada en la IPTC

Con el fin de normalizar el intercambio y comercio de información entre medios, la NIFT incluye una clasificación de materias, para utilizarse como thesaurus, que

aborda las principales áreas temáticas de un medio de comunicación. Se estructura en 17 grandes categorías:

01 CULTURA
02 JUSTICIA
03 CATÁSTROFES Y ACCIDENTES
04 ECONOMÍA
05 EDUCACIÓN
06 MEDIO AMBIENTE
07 SALUD
08 INTERÉS HUMANO
09 TRABAJO
10 VIDA COTIDIANA
11 POLÍTICA
12 RELIGIÓN
13 CIENCIA Y TECNOLOGÍA
14 SOCIEDAD
15 DEPORTE
16 GUERRAS Y CONFLICTO
17 METEOROLOGÍA

A su vez, se desarrolla un segundo nivel con categorías más específicas. Cada epígrafe, tanto en un primer nivel como en el segundo, tiene asignado un código numérico de ocho dígitos. Este sistema de codificación, igual para todos los idiomas, facilita el intercambio de información sin que la lengua suponga un obstáculo.

Al ser una clasificación general y en algunas categorías muy norteamericana, se ha introducido un tercer nivel de descriptores o palabras clave de materias para describir con precisión el contenido del documento y adaptarlo a las necesidades de un medio español. Respetando el código original, se han añadido otros tres dígitos más que nos permite incorporar ese tercer nivel y realizar modificaciones según se actualiza la IPTC original.

- Clasificación de palabras clave onomásticas, de entidades, geográficas e identificadores:

Estas entradas se crearon originalmente como índices y, actualmente, se encuentran almacenadas en bases de datos independientes en forma de registros con información básica de una persona, entidad, etc. Por tanto, estos descriptores desempeñan una doble función: descripción del contenido y alimentación de directorios e índices de la web.

4. Proceso de conversión de un sistema a otros: Fases

El proyecto de conversión de los contenidos de Hércules al sistema editorial se realizó en cinco fases:

1.Establecimiento de equivalencias entre tipologías documentales y Thesaurus

En esta primera fase se cotejaron y se establecieron equivalencias entre las tipologías documentales de Hércules y los géneros usados por Prisacom.

El fin era tener la correspondencia de códigos entre ambos sistemas para que a la hora de procesar automáticamente los contenidos, se asignaran los códigos correctamente, sin perder la información que tenían los documentos en la base de datos de El País.

La labor más ardua fue la de equiparar las entradas de las clasificaciones ya que el thesaurus de El País dispone de aproximadamente 750.000 entradas mientras que Prisacom mantiene unas 10.000 entre IPTC y clasificación de palabras clave (onomásticas, entidades, geográficas e identificadores). Para realizar la conversión de las materias se desarrolló una herramienta con dos apartados, correspondientes a ambos thesaurus. y se fue asignando manualmente cada categoría País a su correspondiente en Prisacom.

Los descriptores onomásticos, de entidades, geográficos e identificadores se procesaron mediante un automatismo que incorporaba las entradas originales de El País a Prisacom y si ya existía en esta última se obviaba.

2. Proceso de importación del contenido en formato texto y transformación a XML

En esta fase se recuperaron las noticias de la edición impresa y cuadernos regionales selectivamente. Primeramente se prescindió de informaciones poco interesantes para una hemeroteca como consultorios, convocatorias, cumpleaños y vida social y después se fue tratando por años ya que el volumen era aproximadamente de 1.200.000 registros.

La transformación de los ficheros en formato texto a XML se llevó a cabo con una herramienta propia que automáticamente captura los textos, les asigna los códigos IPTC y palabras clave marcadas en la anterior fase y los exporta al sistema editorial.

Aquellas noticias que presentaban relaciones en Hércules como las de portada con su desarrollo en la sección correspondiente o las que eran noticias de apoyo de otra principal mantuvieron dicha vinculación.

3. Edición y composición de las portada del diario y las portadillas de Sección

Una vez almacenados los contenidos en el sistema editorial de Prisacom, se procedió a generar automáticamente la composición de las portadas y portadillas de sección del diario para facilitar la consulta de la hemeroteca día a día.

El objetivo era conseguir que los lectores, seleccionando una fecha concreta, les aparecieran las noticias organizadas por las secciones del diario, tal y como

aparecieron publicadas. Y, además, unificar la colección digital ya que los diarios de 2001 y 2002 están editados de esta manera.

4. Incorporación del material multimedia desde el año 2001

Cada diario de la hemeroteca lleva dos imágenes de la portada, una pequeña y otra grande, un archivo en formato PDF de la misma y la foto principal de la primera página.

Las imágenes de portada enlazan con su pdf y nos permiten navegar por otras ediciones. Este proceso se realizó de forma automática, asignando a los archivos su fecha y el tamaño establecido para poder incorporarlo al sistema editorial.

5. Control de Calidad de la Hemeroteca y edición de las 300 mejores portadas de El País

Una vez terminadas las anteriores fases, se llevó a cabo un exhaustivo análisis de la hemeroteca. Se fue comprobando día a día si se habían generado correctamente los diarios, si faltaba alguna portada o portadilla o había alguna sección que no presentaba contenido. El resultado fue satisfactorio ya que se detectaron sólo 345 días que faltaban de 8.248 diarios.

Posteriormente, con el fin de que los diarios de la hemeroteca tengan el mismo diseño seleccionamos las mejores 300 portadas editadas por El País y se realizó la composición manual de las mismas y se editó el material multimedia (fotos e imágenes de portada) y el PDF.

Conclusiones

En este trabajo hemos abordado el proceso de creación de la hemeroteca digital de El País a partir de la información que disponía el diario almacenada en su base de datos. Todavía el proyecto no ha concluido ya que queda por incorporar al sistema de Prisacom los contenidos de suplementos desde el año 2001 hacia atrás y los cuadernos extras que son documentos especiales sobre un evento o un tema concreto.

Además, se va a proceder a la edición y composición de 8.248 portadas para que su diseño sea idéntico al actual y añadir el pdf y material multimedia que acompaña a las primeras páginas de la hemeroteca.

Referencias

1 Extensible Markup Language (XML) http://www.w3.org/XML/
2 News Industry Text Format (NITF) http://www.nitf.org
3 International Press Telecommunications Council (IPTC) http://www.iptc.org