MEI
II, vol. 2
nº 2

Hemeroteca digital de la Biblioteca Nacional: Tres claves de su éxito

Lola Rodríguez Fuentes
Biblioteca Nacional

Recibido el
02-05-2011

Aceptado en
25-07-2011

Resumen
Tras cuatro años desde el lanzamiento de *Hemeroteca digital* en la BNE, y terminada la primera fase del proyecto de digitalización (2008-2010) podemos confirmar su éxito entre nuestros usuarios e investigadores internacionales. En abril de 2011 *Hemeroteca digital* cuenta con 840 títulos y más de 4.000.000 de páginas. Aquí presentamos las conclusiones de la evaluación que hemos llevado a cabo en la que destacan tres puntos clave: contenido, herramienta de búsqueda y servicios de valor añadido.

Palabras clave
OCR; Digitalización de prensa; Servicios en línea; Hemerotecas.

Abstract
Four years after the launching of the *Hemeroteca Digital*, we can confirm its success among our users and international researchers. We have carried out an evaluation to the first stage of the digitization (2008-2010) project. As of April 2011, the *Hemeroteca Digital* amounts to 840 titles, comprising more than 4.000.000 pages. Three are the main points that resulted in the success of the application: contents, searching tools and added value services. These aspects will be described shortly below.

Keywords
OCR; Press digitization; online services; Press archives.

1. Introducción

Tras cuatro años desde la puesta en funcionamiento de *Hemeroteca digital* (marzo 2007) y antes de su integración en *Biblioteca Digital Hispánica*, que supondrá algunos cambios en el proyecto original, hemos realizado una evaluación de esta primera etapa (2008-2010) y los resultados de la misma nos animan a continuar con este proyecto.

Hasta la fecha *Hemeroteca digital* cuenta con 840 títulos, con más de 4.000.000 de páginas. Hemos podido constatar el éxito que este servicio ha tenido entre nuestros usuarios y entre la comunidad investigadora internacional (como muestra el gráfico de consultas).

Del resultado de dicha evaluación podemos señalar tres puntos fuertes como claves del éxito: contenido, herramientas de búsqueda y servicios de valor añadido. A continuación se describen brevemente.

RODRÍGUEZ FUENTES, LOLA. "Hemeroteca digital de la Biblioteca Nacional: tres claves de su éxito". En: *Métodos de información* (MEI), II Época, Vol. 2, nº, 2, 2011, pp. 21-27.
DOI: http://dx.doi.org/10.5557/IIMEI2-N2-021027

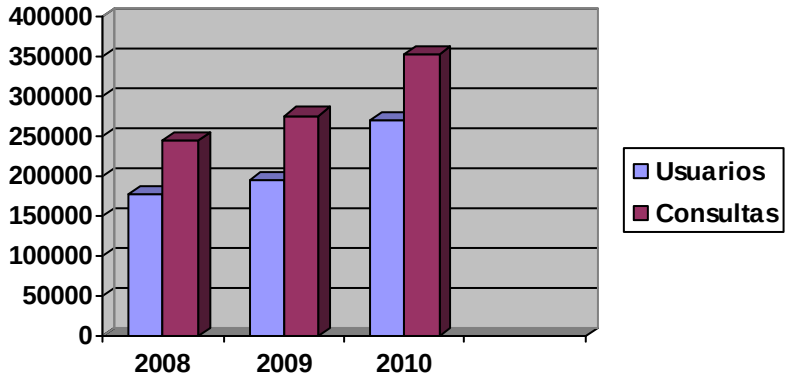

Fig 1. Gráfico comparativo del uso de HD (2008-2010)

2. Contenido

2.1. Selección

Cuando una gran biblioteca patrimonial se enfrenta al reto de digitalizar sus colecciones, el primer paso que tiene que llevar a cabo es establecer que parte de la colección se podrá digitalizar con los recursos disponibles, ante la imposibilidad de tratar las colecciones completas.

La *Biblioteca Nacional* cuenta con unos 8.000 títulos de prensa histórica que podrían digitalizarse y un presupuesto que nos permite realizar la digitalización de unos 800 títulos, esto es un 10% de la colección total. Se impone, por tanto, un proceso de selección, de cómo se lleve a cabo este proceso dependerá en gran medida el éxito o el fracaso del proyecto de digitalización.

En nuestro caso se ha realizado la selección en base a los siguientes factores:

- CONSERVACIÓN: se ha discutido bastante en diferentes foros sobre la conveniencia o no de la digitalización como acción de conservación frente a la más extendida idea de que la microfilmación es el sistema que asegura la pervivencia de los contenidos originales, al tener unos estándares muy arraigados y necesidad para la consulta de equipos poco sofisticados. También se va abriendo camino la idea de que la digitalización apoya al menos los programas de preservación, ya que al no usar los originales para difundir la información, estos no se ven sometidos al deterioro que el uso siempre supone. Siguiendo en esa línea nosotros hemos seleccionado colecciones de prensa fundamentales en muy mal estado, que de otra forma no podrían consultarse, ejemplares únicos de gran interés y publicaciones muy consultadas sometidas a una manipulación y reproducciones parciales constantes que afectan de forma muy negativa a su pervivencia. En este caso el factor predominante de la selección ha sido el estado físico de las publicaciones.


AÑO XXXV.

LA EPOCA.

GAMBETTA Y LA REPUBLICA.

CARTAS DE ITALIA.

Fig. 2. Ejemplo de periódico mutilado

- USO: también hemos querido atender a la demanda de nuestros usuarios, tanto reales como potenciales, por tanto hemos seleccionado los títulos más representativos de nuestra colección, ya que uno de nuestros objetivos es una mayor divulgación de la Biblioteca, mejorando el acceso a nuestras publicaciones más solicitadas y al mismo tiempo buscando la fidelidad de nuestros usuarios. El factor a tener en cuenta ha sido las estadísticas de uso, junto a las demandas reiteradas de nuestros usuarios.

- TIPO DE MATERIAL: además de contar con los títulos más señeros de periódicos españoles históricos, también se ha seleccionado muchas de las principales revistas en cada materia y de las revistas ilustradas de noticias, que presentan una visión muy completa de los acontecimientos políticos, sociales y culturales de su época. Ofreciendo así una amplia oferta representativa de la colección de la BN.

- OTROS: hay otros aspectos que se han tenido en cuenta en el proceso de selección y que complementan el fin perseguido en el proyecto, como son la cooperación entre centros de similares funciones, el respaldo a exposiciones o eventos significativos que han enriquecido la selección con títulos en principio no considerados, sugerencias de entidades o instituciones diversas e incluso sugerencias de investigadores o docentes.

2.2. Metodología

Una vez determinado el objetivo de la selección se comenzó la revisión y el análisis de la colección a reproducir, esta ha sido la tarea más laboriosa del proyecto y ha consumido mucho tiempo y gran parte de los recursos humanos del *Departamento de Seriadas*. La metodología ha sido la siguiente:

- Se ha realizado una base de datos que recoge toda la información de la selección, el análisis y las distintas etapas del *workflow* de cada título.
- Mediante la consulta en el catálogo se ha determinado la existencia o no de duplicados, y sus distintas ubicaciones.
- Se ha revisado página a página todos los ejemplares posibles, seleccionando los más idóneos.
- Se han detectado los problemas existentes en los originales como son: faltas, lagunas, mutilaciones, manchas, roturas con o sin perdida de texto, traspaso de tintas del verso y deterioros varios que se han intentado cubrir con otras colecciones.
- En el caso de no haber podido completar ejemplares o títulos se hace constar en las imágenes y se ha guardado toda esa información para intentar ir completando colecciones mediante la cooperación con otros centros.
- En algunos casos se ha procedido a pequeñas intervenciones de restauración para consolidar el soporte antes de la digitalización, pero en otras ocasiones la perdida de texto ya era irreparable.

2.3. Coordinación intercentros

Se han llevado a cabo algunas acciones de cooperación con otras instituciones con el fin de compartir recursos y completar en la medida de lo posible las colecciones existentes en la BN. Siguiendo esta idea, se han descartado, en principio, todos aquellos títulos que estén digitalizados en otros centros, para lo que se hace una importante labor de investigación en todas las bibliotecas digitales españolas.

En una segunda fase se intentará localizar en otras bibliotecas los ejemplares que faltan en Hemeroteca digital con el fin de completar títulos.

3. Herramienta de búsqueda

3.1. OCR: Búsqueda textual

Por lo que se refiere a herramientas de consulta de prensa en España, históricamente nos hemos encontrado con la carencia de índices o sumarios que nos faciliten la localización de los artículos o de los temas que se quieran consultar. El investigador ha venido reclamando alguna herramienta que le facilite su tarea, sobre todo desde que la investigación en prensa se ha generalizado convirtiéndose en imprescindible para tratar periodos históricos o temas recientes. La posibilidad de aplicar técnicas de OCR a las imágenes digitales ha venido a paliar, en gran medida, la carencia de instrumentos de búsqueda más sofisticados como tesauros o índices, pero también

mucho más costosos y prácticamente inviables en aquellos países con gran cantidad de títulos de prensa.

En todas las bibliotecas que están llevando a cabo importantes proyectos de digitalización (véase *TELplus project*) se está analizando la conveniencia de aplicar técnicas de OCR a periódicos antiguos y la exactitud de los resultados obtenidos. No hay que perder de vista que el OCR siempre será una búsqueda textual y no controlada, por lo que aunque su trascripción fuese exacta seguiría dándonos resultados erróneos (y mucho "ruido" en nombres comunes y fechas).

En nuestro caso concreto habíamos comenzado a experimentar con OCR aplicado a algunos títulos de prensa moderna y a pesar de que los resultados no eran totalmente satisfactorios, si pudimos observar el impacto que tuvo en la investigación y consulta de la prensa. Los títulos más utilizados fueron los que contaban con esta herramienta e incluso se usaba la búsqueda en OCR para averiguar las fechas de los eventos que se trataban de localizar y poder buscar después esa fecha en otros títulos más interesantes para el investigador.

Además cuando los usuarios comenzaron a adiestrarse en el uso de la herramienta fueron capaces de diseñar perfiles de búsqueda que respondían suficientemente a sus consultas minimizando los fallos del reconocimiento de caracteres. A continuación comenzaron a demandar que se digitalizase y se aplicase OCR a toda la prensa histórica que estaba microfilmada para facilitar su consulta.

Con este escenario, ¿cómo actuar al acometer el proyecto de digitalización de prensa histórica?. Evidentemente, nosotros hemos optado por dotar a Hemeroteca digital con la única herramienta de búsqueda de contenido que actualmente podemos ofrecer a nuestros investigadores y a pesar de su inexactitud ellos nos lo han agradecido con una elevada cantidad de consultas y felicitaciones por nuestro trabajo.

El software que estamos empleando es *ABBYY Finereader*, hemos podido apreciar la mejora del producto desde sus primeras versiones a la más actual.

A pesar de no haber realizado un estudio pormenorizado de los resultados obtenidos en la búsqueda mediante OCR, lo que sí hemos hecho ha sido un análisis automático sobre una muestra de 268.696 páginas, pertenecientes a 23 títulos diferentes, en ella los motores de OCR afirman haber reconocido 4.000.867.107 caracteres, consideran caracteres correctos o con un muy bajo umbral de duda a 3.459.095.760 caracteres, por tanto el porcentaje medio correcto de todos ellos es 86,46 %.

Si la media se realiza considerando el acierto medio por página entonces es de 87 %. Estos números indican una relativa buena lectura y que ésta es más o menos homogénea en todas las páginas. Los porcentajes por título están entre un 98% el de mejor resultado de lectura y un 82% el de peor.

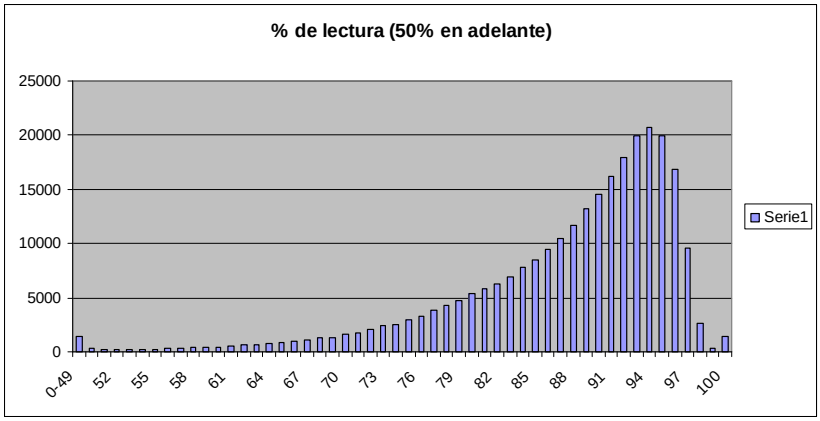

Fig. 3. Análisis automático de la muestra OCR (% de lectura)

Ejemplo: resultado de OCR en un periódico histórico:

Siglo futuro, El (Madrid. 1875). 04/02/1932, n. 7544, página 1. (PDF, 6 páginas, ~2 MB.)
, por- que soy feliz en mi hogar". COMO él señor Unamuno, que de- cía en los pasillos: "Se ha [...] la ley de divorcio no existe; o como si no existiera. Como al isieñor Unamuno, no les importa

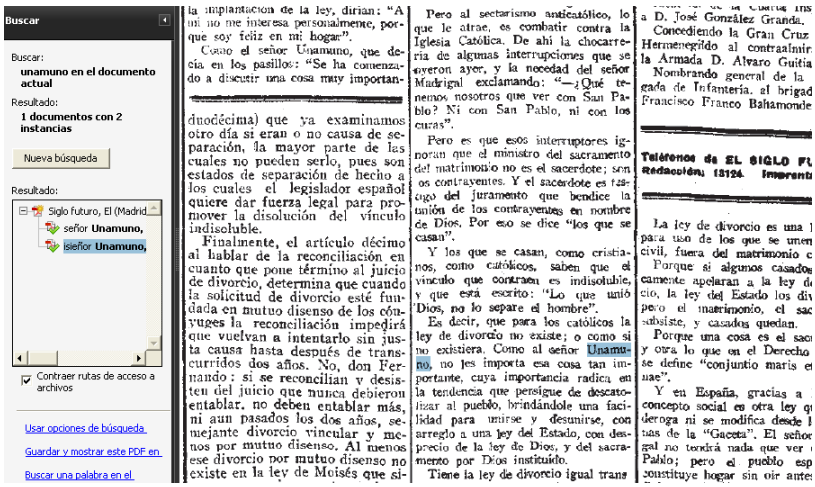

Fig. 4. Ejemplo: resultado de OCR en un periódico histórico

3.2. Posibles soluciones

- Corrección manual con ayuda de un editor de texto, esta medida sería la de resultados más exactos, pero es totalmente inviable para tratar una colección digital del tamaño de Hemeroteca digital, se puede pensar en ella solamente para tratar algún título importante con graves problemas de lectura debido al mal estado del original.
- Volver a aplicar a las imágenes las futuras versiones mejoradas del software, siempre que una muestra representativa demuestre la conveniencia por la mejora en la recuperación.

- Otras alternativas como la corrección cooperativa, con participación de los investigadores que consultan la aplicación, parecen actualmente muy complicadas pero podría ser una opción en el futuro.

4. Servicio de valor añadido

Además de contenido y de facilidades de búsqueda Hemeroteca digital se apoya en una serie de herramientas y servicios que aportan valor añadido a la consulta de prensa histórica española y la convierten en el portal de referencia para los investigadores, como son:

- Catálogo bibliográfico:

Las imágenes digitalizadas cuentan con un enlace al catálogo de la Biblioteca, donde se puede consultar su registro bibliográfico normalizado y todas las colecciones, tanto originales como reproducidas, existentes en la Biblioteca de ese título. También se puede llegar a las imágenes digitalizadas desde el registro bibliográfico del catálogo.

- Descripción de publicaciones:

Las publicaciones digitalizadas tienen una pequeña presentación, donde se explica la importancia de la publicación, su historia y el papel que representó en su época, esto puede ser muy útil para contrastar la información con otros títulos de diferente ideología

- Otras hemerotecas digitales:

Hemeroteca digital cuenta con un directorio donde se recogen los enlaces a las principales hemerotecas digitales españolas con una breve presentación de las mismas y de sus recursos

- Directorio de periódicos electrónicos actuales:

En la Web de la Biblioteca se mantiene un directorio con los enlaces a todos los títulos de prensa española actuales en-línea

- SFX y recursos electrónicos:

A través del gestor de recursos electrónicos SFX podemos localizar los periódicos y revistas electrónicas actuales disponibles en las numerosas bases de datos accesibles en la Web de la Biblioteca.

- Proyectos europeos:

Hemeroteca digital como parte de *Biblioteca Digital Hispánica* forma parte del proyecto *Europeana* y de otros muchos proyectos que se están llevando a cabo en el ámbito de *The European Library*, lo que permitirá en un futuro compartir sus recursos con las principales biblioteca digitales europeas y realizar búsquedas sobre sus colecciones digitales de forma simultanea.