

**DEL PAPEL AL MONITOR:
DIGITALIZACIÓN Y RECURSOS DE INFORMACIÓN
EN INTERNET
PARA LOS USUARIOS DEL ARCHIVO DE REDACCIÓN
DEL DIARIO CLARÍN**

AGUSTÍN MAURIN

Resumen: El presente trabajo reseña el proceso de transformación que se está llevando a cabo en el Archivo de Redacción del Diario Clarín de Buenos Aires, Argentina. La digitalización de diarios y fotografías, así como también la creación de Bases de Datos de negativos e imágenes en formato digital, conforman una respuesta práctica y eficiente a la demandas del diario de mayor circulación de habla hispana. Se detallan, además, los pasos seguidos en el desarrollo de páginas web para dotar a la redacción de herramientas de búsqueda y recuperación de la información tanto a través de una Intranet, como el desarrollo y actualización constante de una completa guía de recursos periodísticos en Internet.

Palabras clave: Digitalización; Imágenes digitales, Digitalización de archivos; Bases de datos; Excalibur; Recursos de información en Internet

Abstract: This paper reviews the transformation process being held by the Clarín Newspaper Archives (Buenos Aires, Argentina). Digitalization of photographs and newspapers, as well as digital image databases are just the starting point to the demands generated by this newspaper - ranked number one among hispanic countries. There is also detailed information on the process by which web pages were developed in order to provide journalists with information search and retrieval tools through an Intranet, and by the development of a complete and up-to-dated guide to journalistic resources in the Internet.

Keywords: Digital Archiving; Digitalization; Newspaper Digitalization; Images Data Bases; Image Digitalization

Jefe de Archivo, Diario Clarín.

Correo electrónico: agustin@maurin.net; amaurin@redaccion.clarin.com.ar

Artículo recibido: 31-10-00. Aceptado: 14-11-00.

INFORMACIÓN, CULTURA Y SOCIEDAD. No. 4 (2001) p. 37-50

©Universidad de Buenos Aires. Facultad de Filosofía y Letras. Instituto de Investigaciones Bibliotecológicas (INIBI), ISSN: 1514-8327.

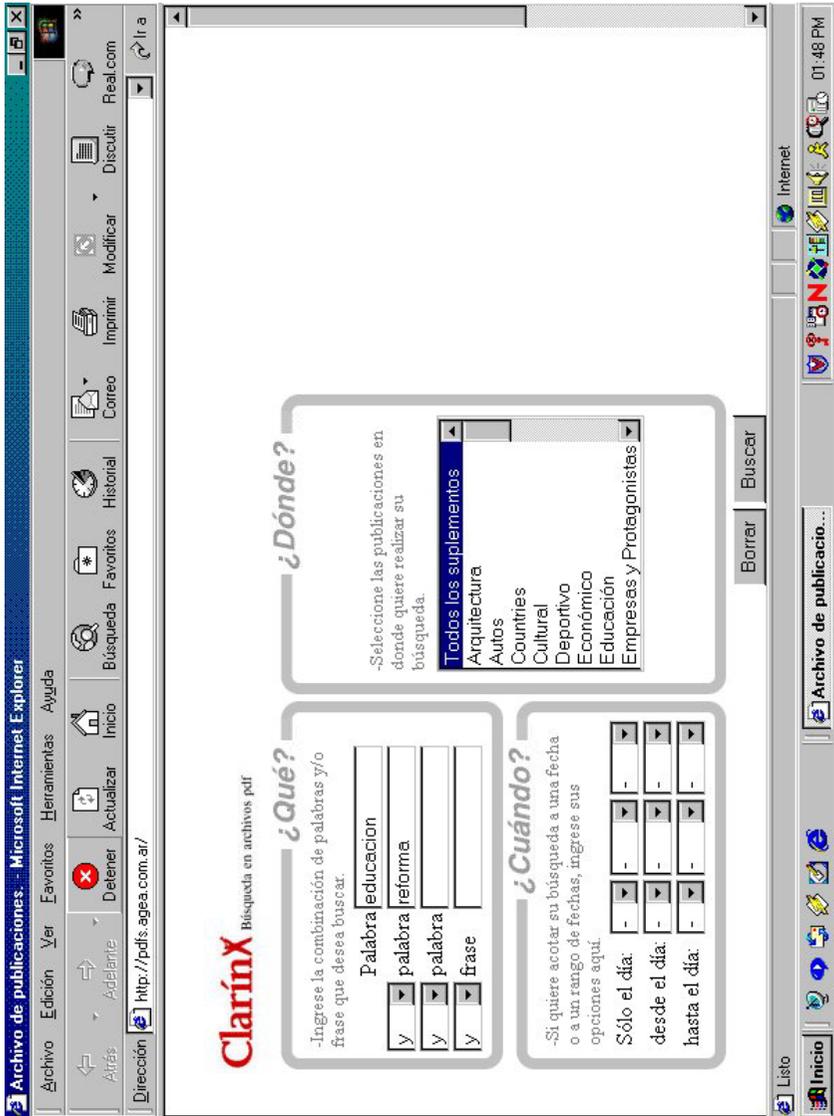


Imagen 1a



Imagen 1b

La era digital ha marcado el fin del milenio con la impronta de una potencia tecnológica jamás conocida hasta hoy por el hombre. La digitalización de papeles y microfilms, la generación de imágenes de alta resolución y su aplicación directa en ediciones gráficas y audiovisuales, y la fabulosa expansión de Internet, que ha generado una impresionante conversión de los formatos conocidos hacia contenidos accesibles vía WWW (World Wide Web), conforman un universo lleno de posibilidades y desafíos. En este contexto, la aplicación de los criterios propios de las Ciencias de la Información, la Bibliotecología y la Documentación, siguen teniendo plena y renovada vigencia.

El Archivo de Redacción de Clarín tiene mas de 50 años de vida. Desde que el diario comenzó a publicarse, se optó por guardar los ejemplares y noticias recortadas de las propias ediciones y de otros medios gráficos. En un principio las ediciones eran de muy pocas hojas, y todo se reducía a encuadernar y conservar. A medida que el diario fue haciéndose más y más voluminoso, y que fueron diferenciándose suplementos y secciones, crecía el número de páginas y el número de periodistas y colaboradores que requerían un servicio organizado y sistemático. En la década del 70 se comenzó a afianzar una idea de archivo que, 20 años más tarde, sería el que encontró al asumir su cargo de Jefe, a fines de 1992, el autor de este artículo. En ese momento, la metodología de trabajo era similar a la que se implementa actualmente, con tres turnos en los que la diferenciación de las tareas está, lógicamente, orientada a la dinámica propia de la redacción. Por la mañana, se recortan los diarios y revistas, se clasifican y luego se archivan hacia el mediodía y la tarde, tratando de atender todas las consultas y, por fin, a la noche se ordena el material que fue retirado en préstamo o consultado in situ. Hace ocho años, el principal y casi único recurso de información eran los sobres de recortes, los tomos encuadernados, los microfilmes, y alguna enciclopedia o diccionario para cotejar términos o datos.

Por otra parte, el material fotográfico se componía, principalmente, de una enorme cantidad de fotografías en blanco y negro guardadas junto a los recortes y se estaban dando los primeros pasos hacia la conformación de un registro computarizado de las producciones fotográficas en negativos que se realizaban en el sector fotografía, ya que hasta ese momento no se había sistematizado el registro, almacenamiento y recuperación temática de los negativos que se producían en las notas periodísticas.

Con el correr del tiempo las agencias internacionales de noticias comenzaron a incrementar sus envíos de fotografías y, además, se inició un proceso que cambiaría radicalmente la concepción del servicio. Las fotografías que antes se enviaban en forma analógica, para que las impresoras locales las reprodujeran para su posterior utilización, pasaron a ser transmitidas en formato digital, con una extensión denominada JPG, que con una gran compresión permite una alta definición con un peso relativamente pequeño en bytes.

Por otra parte el importante desarrollo que tuvieron los sistemas de procesamiento y almacenamiento de imágenes, la aparición de scanners de alta velocidad, discos magnéticos y sofisticados programas de OCR y bases de datos referenciales, hizo posible comenzar a pensar en la posibilidad de transformar la información en papel en otro soporte con imágenes y palabras.

Archivo de recortes de diarios y revistas

Hacia 1995 se comenzó decididamente a trabajar en distintos lineamientos para ir transformando el acervo de recortes y fotos en un archivo digital y, lo que es mas importante, pensando en la posibilidad de que los usuarios tuvieran acceso en línea a esos contenidos y pudieran recuperar lo que necesitaran desde sus puestos de trabajo.

Se formuló un primer objetivo: tratar de digitalizar las páginas del diario tal cual aparecieron a lo largo de los años. El motivo era poder reconstruir completamente los contenidos históricos de las ediciones, asegurando la posibilidad de contar con todos los artículos publicados para su posterior procesamiento. Desde 1946 las ediciones se microfilmaron guardándose copias en negativo y positivo. Alrededor de 1995 se tomó la decisión de dejar de microfilmear para comenzar a digitalizar las páginas y guardarlas en CD-ROM.

Uno de los primeros mitos que se debieron rebatir era que estos soportes tenían una duración limitada y que en poco tiempo los datos se perderían. Los CD-ROM ya tenían mas de 10 años de existencia y con una manipulación adecuada seguían siendo tan útiles como al principio; sin embargo, aún hoy se siguen escuchando argumentos de esta naturaleza.

De todos modos, en pocos meses comenzarán a duplicar las colecciones de CD-ROM, ya que el costo de esta operación, aun con equipos de PC, es tan bajo que no tiene ninguna incidencia presupuestaria, dado que un CD virgen cuesta menos de 2 dólares y se pueden copiar 4 unidades en 1 hora.

La tarea a emprender era entonces comenzar a digitalizar desde 1995 hacia atrás. Se contrato una empresa para hacer esta tarea, pero poniendo como condición esencial que, además, se procesaran con OCR (Reconocimiento Óptico de Caracteres) las páginas y que se indizarán las palabras para poder acceder a las imágenes desde una base de datos referencial. La empresa que resultó elegida realizó esta tarea con scanners de alta velocidad de alimentación automática, y luego de hacer la lectura de los contenidos con una herramienta de OCR en idioma español, conformó la base con el programa Excalibur (Outing, 1999; Grimes, 1999; Chicago Tribune, 2000). Esta tecnología de indización tiene una poderosísima herramienta de indización y recuperación que le permite cotejar cadenas de caracteres e, incluso, brindar resultados por «aproximación». O sea, si no se encuentra exactamente la cadena buscada, nos da aquellos términos que más se parecen a la cadena de caracteres que conforma la palabra.

Esta característica era muy importante por la necesidad de trabajar con palabras que habían sido recogidas a través de OCR en lenguaje natural, lo que permite, en el mejor de los casos, una precisión del 80 - 90 %. Se tenía conciencia de que, por lo menos, una de cada diez palabras estarían incompletas o mal escritas, y que este porcentaje aumentaría a medida que los diarios escaneados fueran más antiguos porque la impresión era menos nítida en aquellos años.

Además, el sistema de edición comenzaba a cambiarse por otro, que también funcionaría bajo redes Novell, en entorno DOS. Por primera vez los textos de las notas se guardarían completamente y habría una opción para realizar búsquedas por rango de tiempo y secciones, combinando hasta tres palabras (opción AND de la lógica booleana) en el mismo artículo. A medida que se comenzó a utilizar el sistema en todas las secciones de la redacción, fue necesario instruir a los redactores sobre las estrategias posibles de búsqueda y la forma más eficiente de encontrar lo que buscaban.

Esta fue la primera experiencia de conformación de una base de datos en formato digital, y aún hoy se sigue utilizando con muy buenos resultados, aunque el entorno DOS no la haga demasiado atractiva ni amigable.

Base de Datos de páginas en PDF accesibles en Intranet

La impresión de las páginas del diario se realiza generando archivos postscripts que permiten generar páginas en un formato sumamente extendido en el mundo editorial, el PDF. Éste permite ver la página con los colores, imágenes y textos absolutamente iguales a la impresión de la página y tiene, además, una gran ventaja sobre la técnica del escaneado, su «peso» en bytes es mucho más bajo. Además, Acrobat ha desarrollado una gran cantidad de herramientas de procesamiento, búsqueda y visualización de estos archivos con lo que su uso se ha extendido con gran rapidez (Adobe Acrobat Reader, 2000)

Se desarrolló un software especialmente pensado para utilizar las facilidades que brinda esta opción, utilizando una interfase visible con un Browser de Internet (MSIE y Netscape) dentro de una Intranet. La búsqueda y previsualización de resultados se realiza dentro del marco HTML, y para la ver la página se activa el Acrobat Reader que brinda la posibilidad de copiar, hacer zoom o imprimir una reproducción de la página en tamaño A4.

La pantalla de BÚSQUEDA es lo más interesante (véase imagen 1a), ya que en ésta se logró reunir todas las opciones combinatorias de la lógica de Boole, para acotar al máximo el rango de tiempo y las combinaciones de palabras. Como es común a todas las bases de datos en Internet, los operadores Y (AND), O (OR), y NO (NOT), permiten combinar tres palabras y una frase.

Se puso especial atención a la posibilidad de dirigir la búsqueda a la sección principal del diario o a cualquiera de los suplementos que se publican por separado, incluso Clasificados y el Diario Deportivo OLE. Los resultados

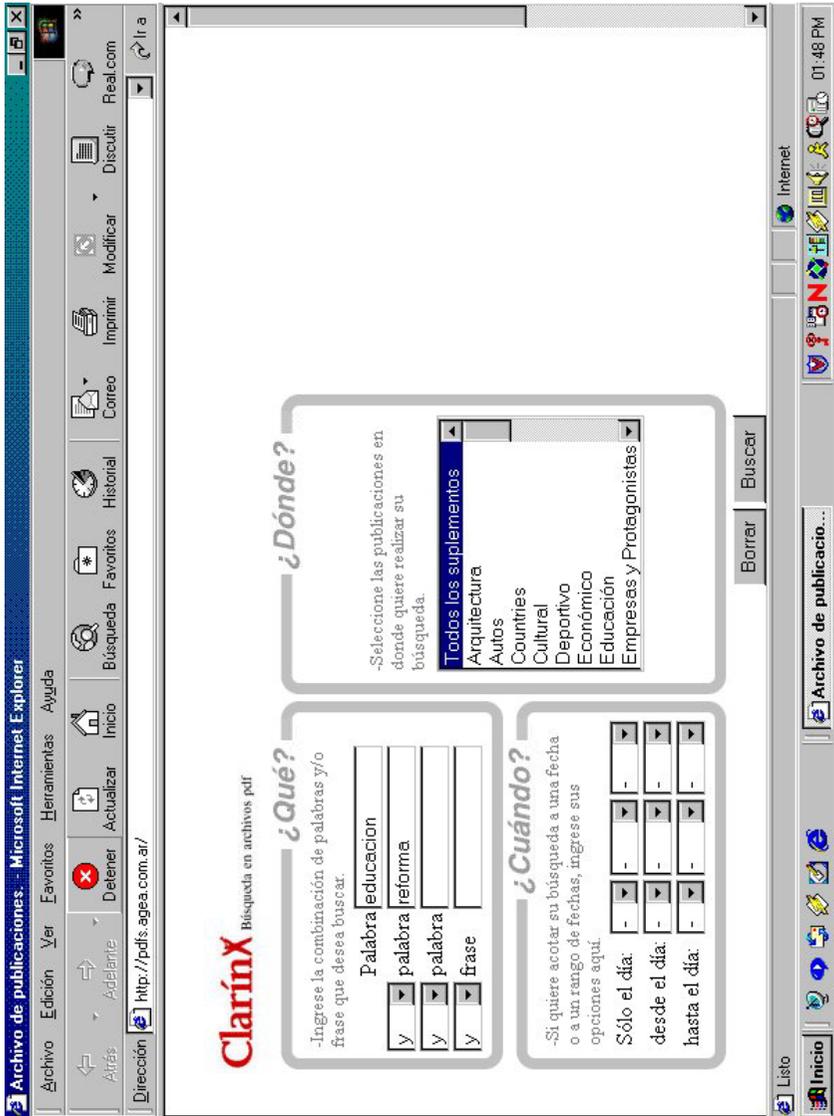


Imagen 1a



Imagen 1b

de la búsqueda se listan en una página que tiene links directos a las páginas del diario que responden a la consulta y se visualizan con un solo clic. (imagen 1b)

Archivo Fotográfico

El archivo de Clarín cuenta con la colección de fotografías en papel blanco y negro más importante del país. La cantidad y variedad de personajes y temas hace imposible saber con certeza su número, pero este alcanza cientos de miles. Hasta no hace muchos años, el proceso editorial requería de la utilización directa de la copia en papel, que se enviaba a Fotograbado, para volver y guardarse nuevamente en sus respectivos sobres y cajones. El desarrollo de sistemas editoriales computarizados, los scanners de alta resolución y la configuración en redes de la interacción página, texto e imagen permitió que, de a poco, se fuera prescindiendo del elemento foto en su soporte tradicional, utilizándose directamente los negativos. Por otra parte, la revolución que llevaron a cabo las agencias internacionales de noticias cuando comenzaron a enviar sus imágenes sólo en formato digital transformó completamente la forma de trabajo en el área de edición y, por lo tanto, se hizo necesario incorporar rutinas de recepción, selección y registro de fotografías en un archivo digital de imágenes interactivo, donde los archiveros registrarán cada una de las fotos seleccionadas por los editores y que, al mismo tiempo, permitiera, en línea, consultar la base de datos, seleccionar y enviar a diagramación.

Diariamente los fotógrafos realizan un sinnúmero de notas y producciones fotográficas. La enorme cantidad de negativos que se genera debía ser conservada y ordenada para facilitar su uso posterior. Se desarrolló un software a medida que permite cargar, en una sola pantalla, todos los datos relevantes para identificar lo atinente a la producción fotográfica y que da la posibilidad de registrar datos no convencionales, tales como actitudes. (véase imagen 2)

Además, se puso especial énfasis en la modalidad de recuperación, para que se pueda acotar la búsqueda con una gran variedad de combinaciones lógicas. De esta manera, cada una de las notas realizadas por los reporteros gráficos se transforma en varios rollos de negativos, que convenientemente acondicionados, se pueden guardar en sobres que se ordenan con una numeración sucesiva, dentro de la fecha de registro, en la base de datos.

Uno de los campos se dedica a consignar esta fecha que funciona como una signatura topográfica para la posterior ubicación, ya que los sobres de negativos se guardan en cajas plásticas ordenadas en estanterías móviles. Los archiveros procesan continuamente el material que se va produciendo y, además, atienden las consultas de los editores fotográficos para dar con el negativo que pueda ser adecuado para ilustrar determinada nota, mediante una pantalla de búsqueda que permite combinar varias opciones. (véase imagen 3)

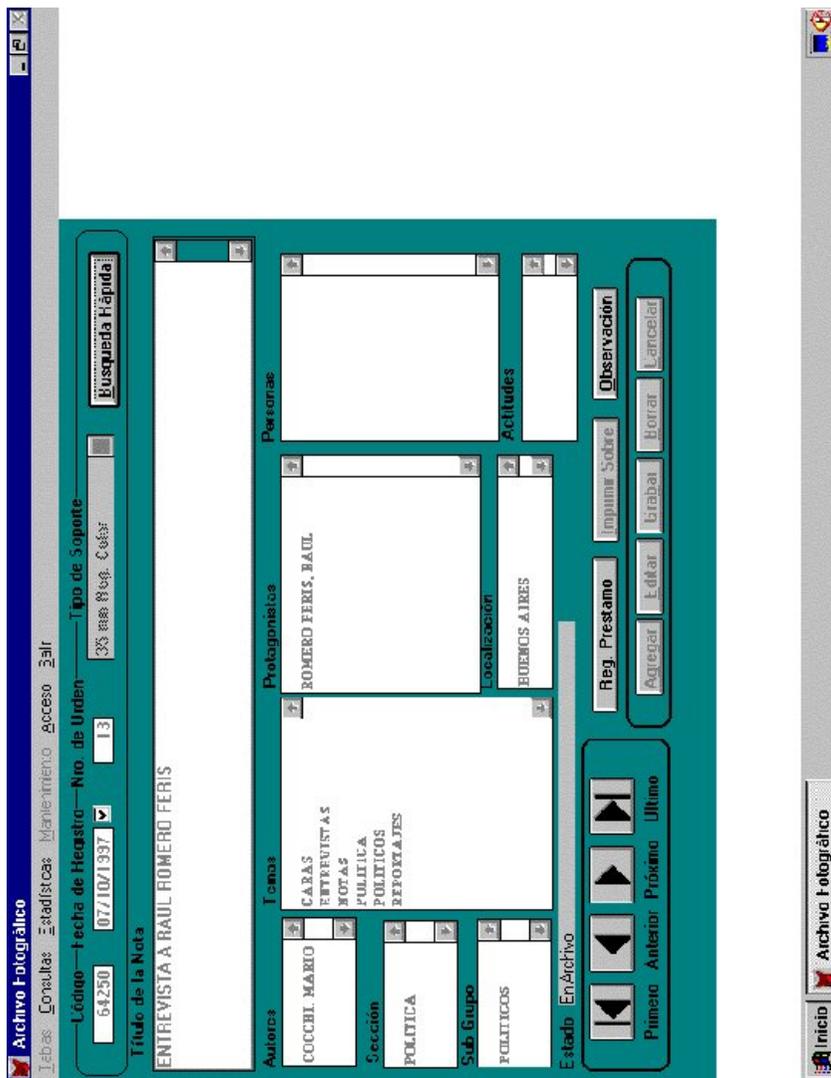


Imagen 2

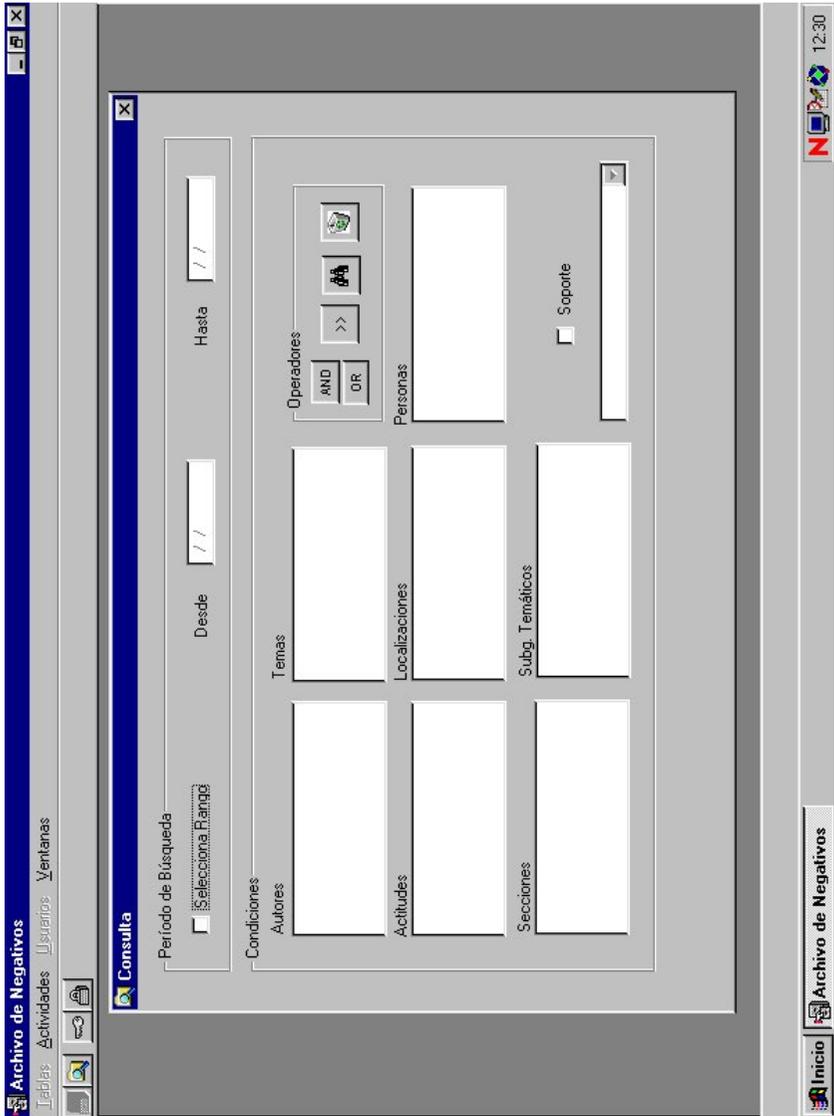


Imagen 3

Fotos en Formato Digital

Hasta 1997, Associated Press, AFP, Reuter, ANSA y todas las agencias nacionales e internacionales, enviaban sus fotografías en forma analógica. O sea, tres capas de rojo, azul y amarillo conformaban el producto final. Estas imágenes se trasmitían para ser reproducidas en forma remota, imprimiéndose por partes o conjuntamente en impresoras que necesitaban de papel especial. Los editores fotográficos seleccionaban estas reproducciones y muchas eran guardadas en el archivo junto con las otras fotos en papel. Con el tiempo, estas reproducciones se tornaban amarillentas y su utilización era altamente improbable. Las tecnologías de transmisión de datos y la gran capacidad de almacenamiento y velocidad de procesamiento, hoy posibilitan la transmisión de las fotografías con una alta definición desde lugares remotos y la recepción, en forma continua, por estaciones receptoras, donde la imagen se ve sin que sea necesario imprimir en papel ni una sola de estas fotos.

En el Archivo Fotográfico de Clarín se desarrolló un software especialmente dedicado a la tarea de ver imágenes, cargar pantallas de descripción con una gran variedad de campos y poder luego recuperar según distintos criterios las fotografías que mejor respondieran a determinada necesidad editorial. El proceso consta, básicamente, de tres pasos: 1) diariamente, se carga una hoja de entrada para cada fotografía donde se registran los datos relevantes; 2) se procede a la indización de la base y 3) se almacenan las imágenes en CD-ROMs que se ubican en un “juke box” donde se hace referencia al registro, para posibilitar una posterior ubicación de la imagen cuando ésta tenga que ser usada. La búsqueda es sencilla y práctica, permitiendo la utilización de distintos criterios y con la visualización de la foto en un recuadro de baja resolución (véase imagen 4). El número de fotografías almacenadas hasta el presente supera las 300.000.

Internet

A partir de la irrupción de la WWW (World Wide Web) en el mundo de la información, los medios de comunicación comenzaron a volcar sus contenidos en la red, presentándose los años 1996 y 1997 como los más prolíficos en este sentido. Los diarios y revistas, y en menor medida emisoras radiales y televisivas, pusieron sus contenidos en sitios y portales, al tiempo que Internet se convertía en un nuevo punto de referencia (Maurin, Olmedo y Susco, 2000). Además, muchas de las fuentes de información en soporte papel, como el caso de la Enciclopedia Británica y de otras bases de datos que estaban en CD-ROM o que ofrecían acceso vía MODEM a costos altísimos, comenzaron a mudar sus contenidos a la WWW (World Wide Web), transformando completamente sus criterios de comercialización y, lo que es más importante, brindando servicios mas económicos, ágiles y actualizados a sus usuarios. Por otra parte, empresas



Imagen 4

privadas y organismos oficiales, fueron incorporando sus bases de datos a la red, conformando un universo vasto y complejo, de esta forma, Internet se convertía en un fabuloso recurso de información. Al mismo tiempo, se hacía evidente la necesidad de facilitar a la redacción del diario lineamientos y guías que permitieran el máximo aprovechamiento de los sitios disponibles.

Desde principios de 1999, se diseñaron páginas en las que se ordenan las temáticas propias de las distintas secciones y suplementos del diario, guías de buscadores y meta-buscadores, listados de diarios, revistas y archivos en línea, y una variada gama de bases de datos científicos, periodísticos, legislativos y estadísticos. La red de redes es el ámbito en donde se ha hecho más patente esta nueva realidad mediante la cual, lo que antes se leía en el papel hoy se lee, copia, guarda o imprime desde una pantalla. Ya hace más de un año que, periódicamente, se dictan cursos de capacitación en la utilización de la GUÍA DE RECURSOS DE INFORMACIÓN PERIODÍSTICA EN INTERNET y, por medio del correo electrónico, se informa a toda la redacción sobre actualizaciones y novedades. La guía de recursos que se diseñó y se mantiene es la página de inicio de los navegadores instalados en las 60 estaciones de acceso a Internet diseminadas por toda la redacción (vease imagen 5). Las estadísticas que se generan automáticamente en el servidor donde está alojada la página, nos permiten asegurar que, con el paso del tiempo, su uso se ha multiplicado en forma geométrica. Baste decir que en el trimestre julio-agosto-septiembre de 1999 la cantidad de accesos era de 3.881, mientras que en el 2000 los “hits” registrados alcanzaron para el mismo período la suma de 17.095.

Sumario

El proceso de digitalización del archivo de Clarín ha permitido aproximarse a una serie de conclusiones preliminares; ellas son, en líneas generales, las siguientes:

- 1) Transferencia paulatina (y a veces vertiginosa) del mundo de lo impreso al ámbito digital, condicionada por los comportamientos de los usuarios y por la aparición, siempre renovada e intensa, de las nuevas tecnologías de la información.
- 2) Fue posible rebatir la afirmación de que el soporte CD-ROM tenía una duración limitada, dado que, en este caso, los mismos superaban los 10 años de existencia conservando su utilidad.
- 3) Este caso ilustra la concreción de una nueva modalidad de acceso a la información, donde el usuario no tiene que ir a buscarla al repositorio sino que puede tenerla en su mesa de trabajo.
- 4) Las nuevas posibilidades de manejo de la información demandan que los profesionales deban encarar programas de capacitación del usuario y compilar herramientas que faciliten su trabajo.



Imagen 5

En síntesis, las experiencias de digitalización pueden variar de uno a otro ambiente de trabajo, pero si responden a las necesidades de la comunidad a la cual sirven implican un mejoramiento en los niveles de desempeño de las mismas. Es preciso, antes de iniciar este tipo de proceso, tener muy en claro el por qué, para qué y sobre todo, quiénes se van a beneficiar.

Bibliografía

Adobe Acrobat Reader. <<http://www.adobe.com/epaper/main.html>> [Consulta: octubre 2000]

Chicago Tribune Select Excalibur to provide Search Solution for Internet. <http://www.excalib.co.za/News/pressrelease7_ChicagoTribune.htm> [Consulta: octubre 2000]

Grimes, Brad. 1999. Copley wilds Excalibur. *TechNews*. Vol. 5, no. 1. <<http://www.naa.org/technews/tn990102/copley.html>> [Consulta: octubre 2000]

Maurin, Agustín; María Inés Olmedo y Jessica Susco. 2000. Medios de comunicación & Periodismo: primera guía comentada de sitios en Internet. Buenos Aires: Alfagrama. 248 p. (Primeras guías comentadas de sitios en Internet)

Outing, Steve. 1999. The Business Case for Digitizing Oldest Archives. *E&P Online*. October 27. <<http://www.mediainfo.com/ephome/news/newshtm/stop/st102799.htm>> [Consulta: octubre 2000]