

Buenas prácticas para la digitalización de documentos destinados al repositorio Digital.CSIC

Isabel Bernal, Juan Román / Oficina Técnica Digital.CSIC

Carolina Santamarina / Unidad de Recursos de Información Científica para la Investigación, URICI-CSIC

04/06/2013

Índice

1. Introducción	3. Copyright
2. Pautas	4. Ejemplos erróneos
a) Tipología de documentos	5. Glosario
b) Formato de imágenes	
c) Color – b/n y resolución	
d) Programas	
e) Tamaño	
f) Recortar/enderezar	
g) Denominación de los ficheros	
h) Compresión	
i) Otras mejoras	
j) Cómo hacer el pdf multipágina (combinar)	
k) Metadato sobre el escaneo	
l) ¿Subir las dos copias? jpg y pdf	

1. Introducción

Actualmente la mayoría de las bibliotecas, archivos o instituciones culturales disponen de escáneres de mesa para uso interno del personal o de autoservicio, como sustituto de la fotocopia. Estos escáneres, enfocados a un uso casi doméstico, permiten digitalizar documentos con el objetivo único de la difusión y/o transmisión.

En esta guía ofreceremos unas buenas prácticas para la digitalización de documentos en el repositorio institucional Digital.CSIC: **digitalización a nivel usuario**, en contraposición a la digitalización con fines de preservación según el Plan Director de digitalización para los fondos del CSIC¹ (**proyectos de digitalización**).

Estos documentos deberían cumplir con unos estándares mínimos de calidad, para que su lectura y manejo sea amigable y tengan un aspecto regular ya que son la carta de presentación de la institución.

Existen muchos modelos de escáneres pero todos van a tener un driver asociado que va a permitir visualizar la imagen y realizar ciertas funciones básicas, como puede ser el recortado, el enderezado, etc...

Pero además, uno se puede ayudar de alguno de los softwares gratuitos como **Irfanview** o softwares que pone a disposición el CSIC para todos los usuarios, como **XnView**, o simplemente los que ofrece Microsoft Office: **Document Imaging** (tiff), **Document Scanning** o **Picture Manager**.

La casuística es tan variada que sólo vamos a dar unas pautas generales para guiar levemente al usuario.

¹ <http://bibliotecas.csic.es/proyectos-de-digitalizacion.-simurg>

2. Pautas

Cómo mínimo una reproducción digital debería responder a estos criterios:

- un escaneo = una página
- las imágenes deben escanearse lo más rectas posible o enderezarlas posteriormente
- deben recortarse, a ser posible todas a la misma medida. **Esto es fundamental**
- deben nombrarse adecuadamente y de forma consistente

Finalmente, con el **Adobe Acrobat Pro** (el CSIC tiene licencia institucional y está en todos los ordenadores) o con otro programa gratuito para la realización de pdf, se deben combinar las imágenes en formato jpg para formar un pdf multipágina que será el que se suba al repositorio.

a) Tipología de documentos

La mayor parte de ítems disponibles en Digital.CSIC son nativos digitales. Sin embargo, pueden darse casos específicos en que sea interesante digitalizar material analógico institucional para darle difusión.

Entre las tipologías de documentos de producción científica institucional que se inscriben en estos casos incluimos:

- Material retrospectivo (artículos, capítulos de libros,...)
- Mapas
- Fotografías
- Otros: póster, dibujos...

Es fundamental recordar que **Simurg** es la plataforma adecuada para albergar y dar difusión a *colecciones patrimoniales digitalizadas* según el Plan Director para la digitalización de Fondos del CSIC.

b) Formato de imágenes

- **GIF** (Graphics Interchange Format): se creó con la finalidad de obtener archivos de tamaño pequeño. Es adecuado para guardar imágenes no fotográficas como logos, dibujos,... Guarda imágenes de 8 bits (256 colores como máximo).
- **JPEG** (Joint Photographic Experts Group): es uno de los formatos más conocidos para la compresión de fotografías digitales usado en todas las cámaras digitales y escáneres. Soporta 24 bits.
- **JPEG2000**: puede trabajar con niveles de compresión mayores que en el caso anterior sin que de un aspecto borroso.
- **PNG** (Portable Network Graphics): apareció para solventar las deficiencias del formato GIF y permite almacenar imágenes con una mayor profundidad de color. Está basado en un algoritmo de compresión sin pérdida. Puede llegar a soportar hasta 24 bits.
- **TIFF** (Tagged Image File Format): archivo estándar para guardar imágenes de alta calidad y muy usado en la impresión de trabajos que utilizan imágenes. De uso común en escáneres. Soporta 48 bits.
- **PDF** (Portable Document Format): formato de almacenamiento de documentos digitales. Guarda con toda precisión el diseño del archivo incluyendo texto, elementos multimedia, hipertexto,...

c) Color – b/n y resolución

Es un error común elegir color cuando no es necesario, incluso cuando éste es mínimo (en títulos, apartados...). Si una fotocopia a veces en b/n es aceptable, lo mismo tenemos que pensar del escaneado.

Lógicamente si el documento incluye gráficos, diagramas,... el color es necesario.

¿Escala de grises? Puede ser importante para fotografías en b/n o en escala de grises pero añade algunas complicaciones del escaneo en color aunque ocupando menos tamaño.

Resolución-color: se deben de escanear a la resolución más baja aunque el fichero superará en tamaño al escaneado en b/n. Es aceptable 200 dpi² (dots per inch, puntos por pulgada).


Resolución-B/N: 300 dpi.

En general recomendamos 300dpi. La resolución y el formato de salida determinan el tamaño del fichero.

d) Programas

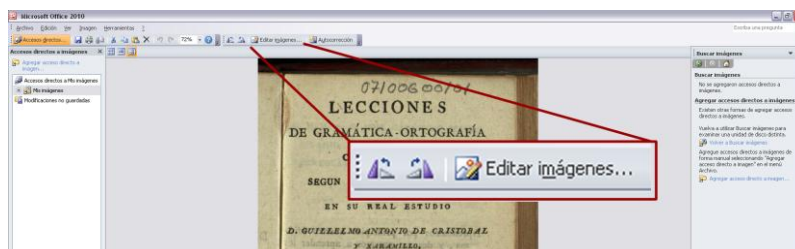
Picture Manager (jpg)³

Inicio → Programas → Microsoft Office → Herramientas de Microsoft Office [año] →

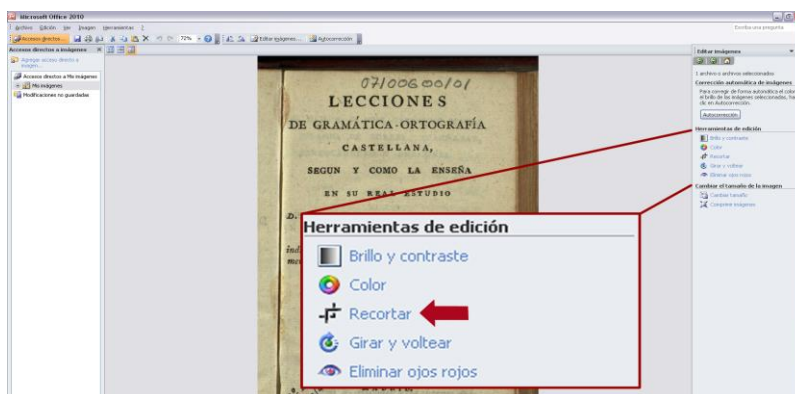
 Microsoft Office Picture Manager

Se abre la imagen con el “Picture Manager”

- se pulsa en **Editar imágenes**:



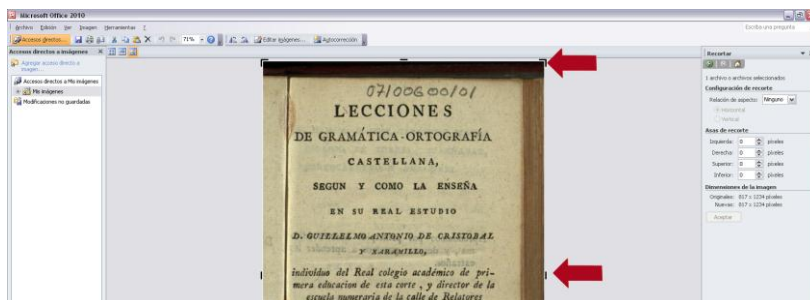
- Se utiliza por ejemplo, la función **Recortar** en el panel de la derecha para eliminar las zonas oscuras que sobran:



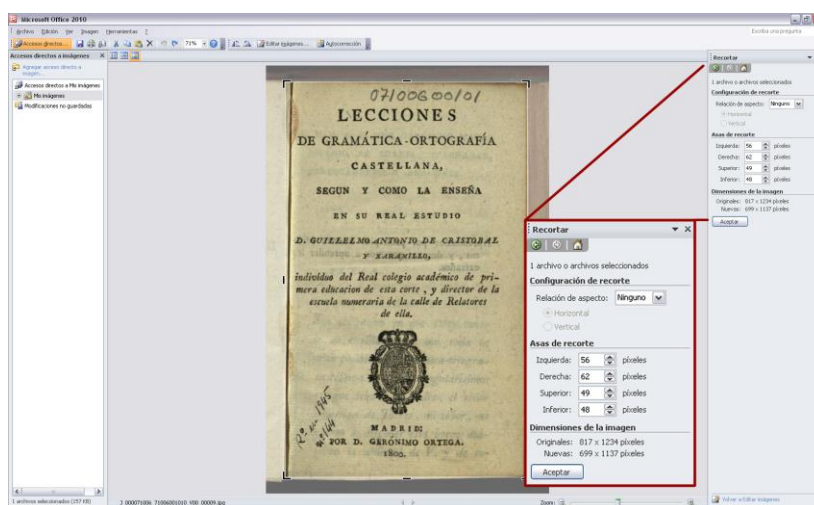
² Define la calidad del escaneo.

³ Estos pasos son similares en cualquier otro programa de edición de imágenes.

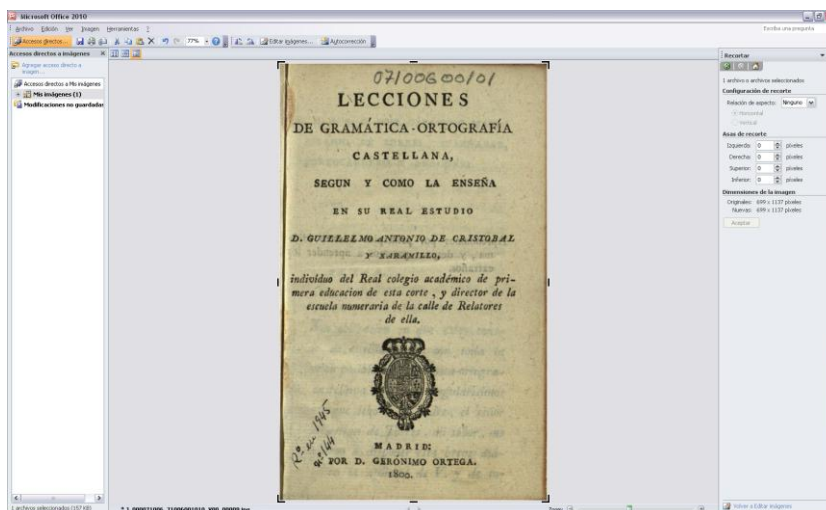
- se cogen las esquinitas y se mueven hasta donde se desea:



- en el panel de la derecha, se pueden ver las dimensiones resultantes finales. Pueden servir de orientación para realizar la siguiente página de forma similar



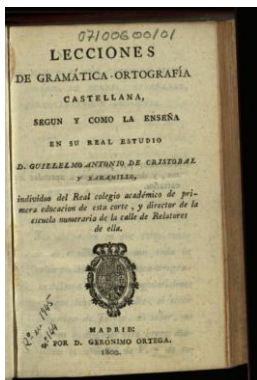
- se guarda la imagen



e) Tamaño

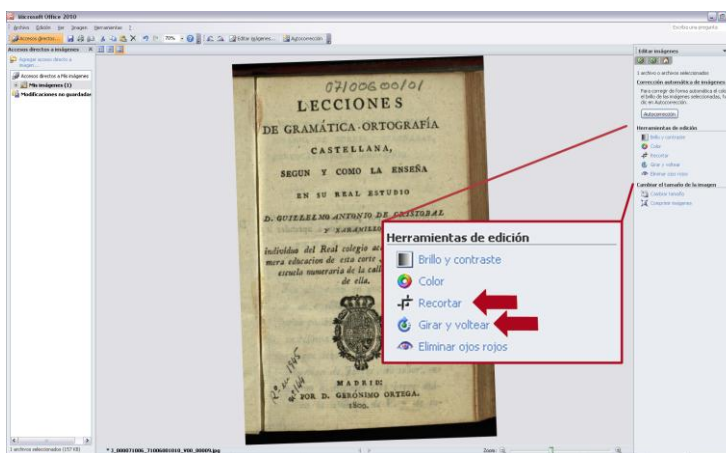
Como norma general podemos tener en mente siempre el tamaño del DIN A4: 21x29.7 cms / 2480x3508 píxeles.

f) Recortar/enderezar



Al escanear una imagen puede que ésta no aparezca recta y que sea necesario recortar para evitar zonas oscuras.

Para solucionar esto podemos utilizar las funciones de **Recortar** y **Girar y voltear** que aparecen en todos programas de edición de imágenes. Por ejemplo en **Picture Manager** aparecen en la parte derecha:



g) Denominación de los ficheros

Hay que tener especial cuidado a la hora de nombrar a las imágenes/escaneos y al pdf final para facilitar una buena gestión de los recursos electrónicos y su accesibilidad:

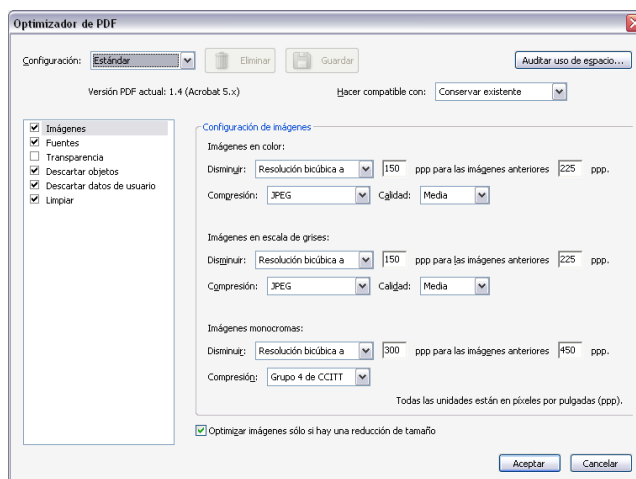
- Imágenes/escaneos: este tema es importante para ordenarlas a la hora de elaborar el pdf multipágina
- Pdf final: evitar documento1.pdf,...; por ejemplo nombrar con alguna palabra clave del título y el apellido del primer autor → mapa_toledo_sXIII_diaz.pdf
- Buenas prácticas en Digital.CSIC⁴:
 - o Elegir nombres que sean suficientemente descriptivos del contenido (por ejemplo, usar parte de la referencia bibliográfica)
 - o Evitar llamar a los ficheros por un número, ya que no es suficientemente descriptivo del contexto del registro
 - o Evitar el uso de caracteres como , . : ; / () \$ & | [] * < > “ ¿
 - o Usar el guión bajo (_) mejor que espacios en blanco entre palabras
 - o No superar los 25 caracteres
 - o Para registros que tienen varias versiones, es aconsejable diferenciar unas versiones de otras indicando v01, v02 etc, en vez de update, nuevo etc. Una excepción resulta cuando denominamos a la versión definitiva, en cuyo caso puede llamarse FINAL. Para estos casos, se aconseja ser consistente en el nombre de los distintos ficheros para guardar una coherencia interna.

⁴ <http://digital.csic.es/faqs/#faq26>

h) Compresión

Hay que evitar en la medida de lo posible los archivos pesados ya que todo lo que supere los 50 MB es, en general, demasiado para su descarga en Internet. Si fuera necesario, se puede proceder a la compresión de éstos teniendo siempre en mente que no debe de producir una pérdida de calidad excesiva.

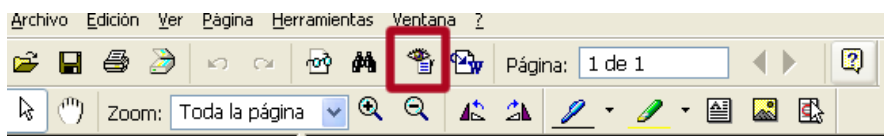
Por ejemplo, el **Adobe Acrobat Pro** ofrece la posibilidad posterior de comprimir el pdf: **Avanzadas → Optimizador de PDF**



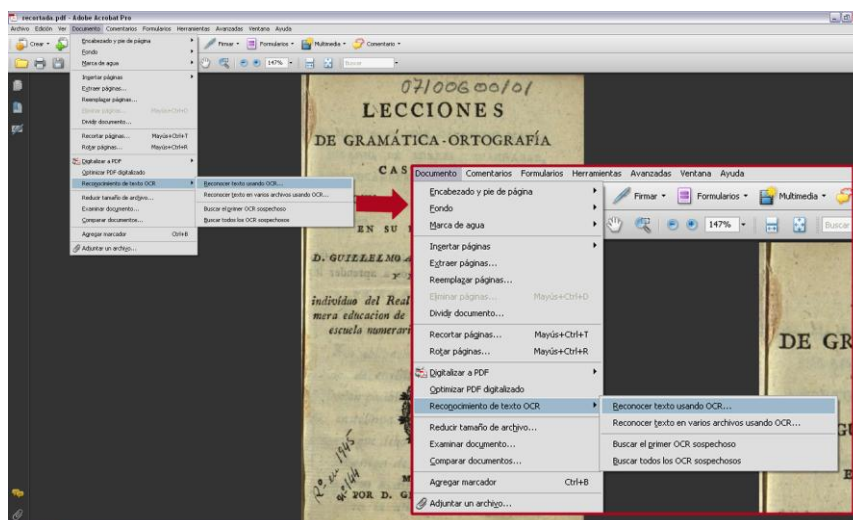
i) Otras mejoras

Se puede realizar un **OCR** (Optical Character Recognition) sobre el documento, aunque esto supone la pérdida del formato original del documento. Esto puede hacerse con el **Document Imaging** o directamente con **Adobe Acrobat Pro** en todo el pdf multipágina.

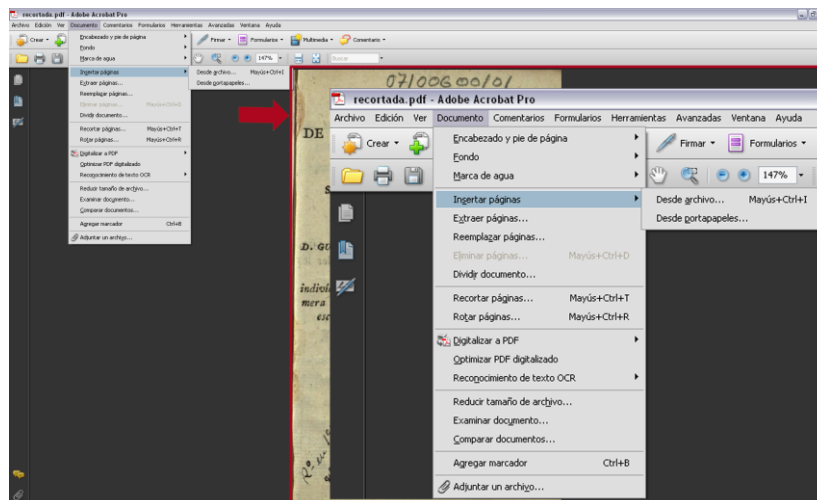
Document Imaging:



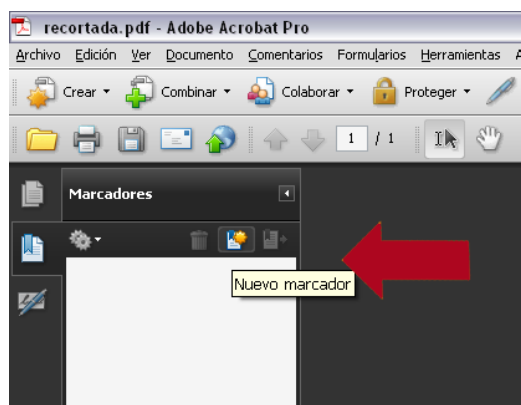
Adobe Acrobat Pro:



También se le puede insertar una portada que de uniformidad a todos los documentos:



Realizar marcadores al pdf:

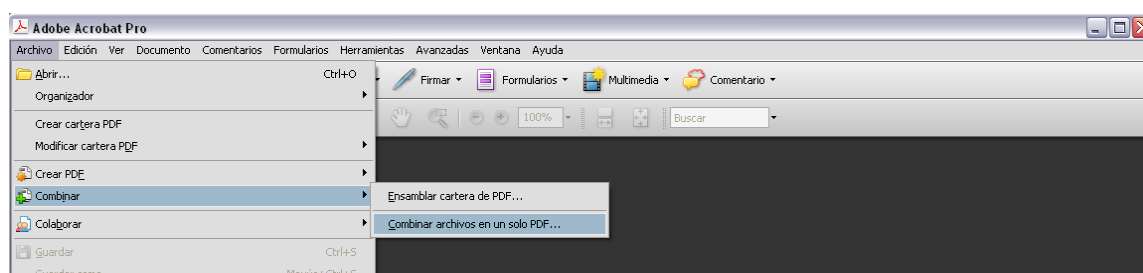


Se pueden realizar muchas mejoras que por supuesto llevan tiempo pero que incrementan la calidad de las presentaciones. Ya dependerá de lo que cada usuario quiera hacer, del tiempo que quiera emplear y de la habilidad en el manejo de cada herramienta.

j) Cómo hacer el pdf multipágina (combinar)

El fichero resultante se debe convertir en otro fichero con formato **PDF** que es uno de los formatos más extendidos para el intercambio de documentos. Como cada escaneo es una imagen es necesario combinar todas las imágenes para generar un único pdf (si el documento resultante es únicamente una página se convertirá a pdf).

Adobe Acrobat Pro: Archivo → Combinar → Combinar archivos en un solo PDF



k) Metadato sobre el escaneo

Para los trabajos escaneados en el campo **dc.description.provenance** puede incluirse la siguiente información separado por .--:

- Fecha del escaneado
- Modelo de escáner utilizado
- Autor del escaneado (biblioteca)
- Localización del original
- Si hay restricción en el uso del material escaneado

Por ejemplo:

2013-05-25.-- Ricoh 502.-- Biblioteca URICL.-- Biblioteca URICL, estantería 502.-- Sólo se permite el uso con fines educativos y de investigación.

l) ¿Subir las dos copias?: jpg y pdf

Se podrán adjuntar al repositorio tanto el pdf final resultante como las distintas imágenes (jpg) siempre y cuando no sean muy numerosas (aunque podrán ir todas en un archivo comprimido .zip, .rar, etc...).


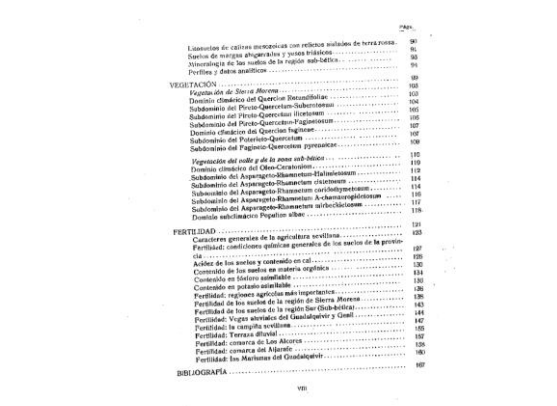
3. Copyright

Escanear una obra sujeta a derechos de autor y subir el resultado al repositorio entraña varios actos de explotación y por tanto se requiere la autorización expresa de los titulares de los derechos (los autores o los editores) previamente.

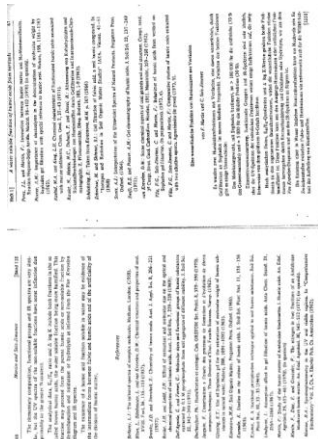
A la hora de considerar el escaneo de una obra, son condicionantes muy relevantes si es de carácter venal, si ha habido transferencia de derechos de explotación a terceros y si el contrato de transferencia sigue vigente.

Se recomienda encarecidamente hacer las gestiones pertinentes y enviar a la Oficina Técnica de Digital.CSIC una copia de la autorización expresa.

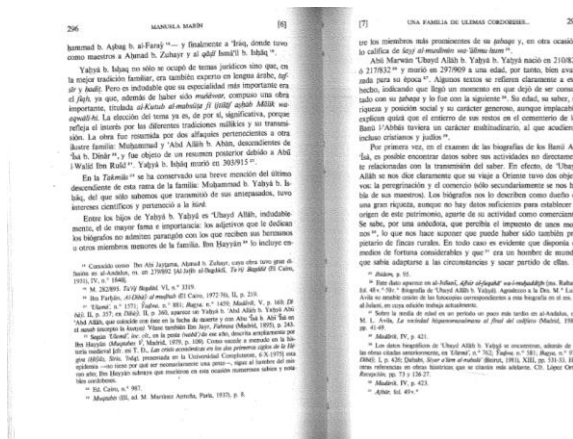
4. Ejemplos de malas prácticas

<p>Imagen no recortada correctamente. Márgenes “negros” <u>Solución:</u> editar imagen (recortar)</p>	<p>Página no enderezada. <u>Solución:</u> editar imagen (girar)</p>
	

Orientación del texto incorrecta.
Solución: editar imagen (girar)



Texto "inclinado".
Solución: una página/un escaneo y editar imagen si es necesario (girar)



Distinto tamaño en las diferentes imágenes/escaneos.
Solución: editar imagen (recortar todas iguales)

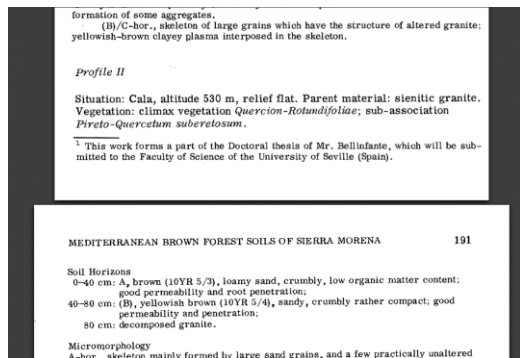
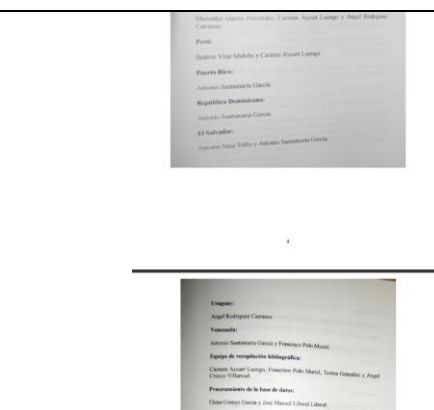
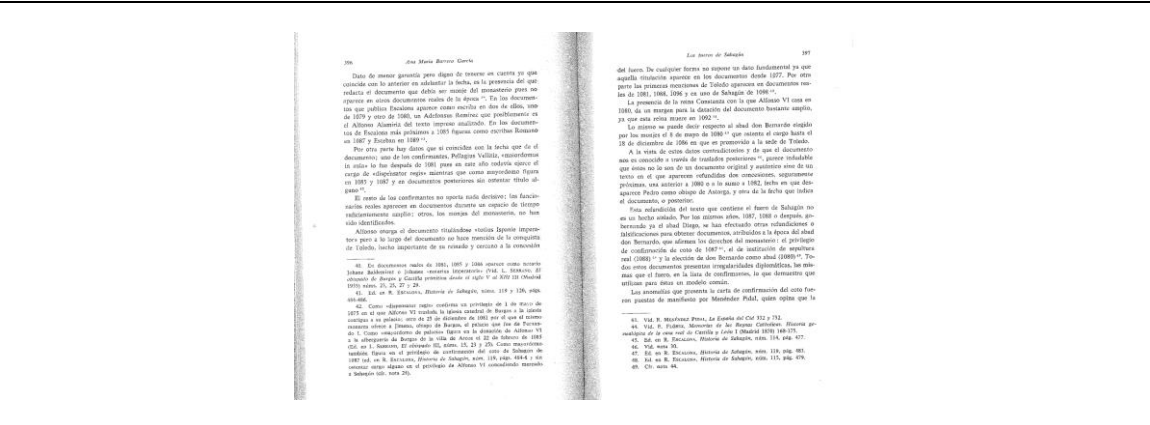


Imagen pegada en un documento Word.
Solución: convertir directamente las imágenes en pdf multipágina



Sombras.
Solución: editar imagen (recortar/eliminar)



5. Glosario

Adobe Acrobat Pro: programa informático de Adobe Systems diseñado para visualizar, crear y modificar archivos con el formato PDF.

Bit (Binary digit): dígito del sistema de numeración binario que es la unidad mínima de información empleada en informática.

Compresión (de archivos): reducción del tamaño de un fichero para evitar un peso excesivo.

Copyright: conjunto de normas jurídicas y principios que regulan los derechos morales y patrimoniales que la ley concede a los autores.

DPI (dots per inch): puntos por pulgada. Unidad de medida para resoluciones de impresión, número de punto individuales de tinta que una impresora o tóner puede producir en un espacio lineal de una pulgada.

Document Imaging: software de Microsoft que permite escanear y trabajar con el document escaneado.

Formato: disposición para formalizar los datos de un documento (procesador de texto, hojas de cálculo, base de datos,...).

GIF (Graphics Interchange Format): creado con la finalidad de obtener archivos de tamaño pequeño. Es adecuado para guardar imágenes no fotográficas como logos, dibujos,... Guarda imágenes de 8 bits (256 colores como máximo).

Irfanview: software gratuito que permite ver, convertir, optimizar y escanear imágenes (entre otras funciones).

JPEG (Joint Photographic Experts Group): es uno de los formatos más conocidos para la compresión de fotografías digitales usado en todas las cámaras digitales y escáneres. Soporta 24 bits.

JPEG2000: puede trabajar con niveles de compresión mayores que en el caso anterior sin que de un aspecto borroso.

Marcador (en pdf): especie de “índice hipertextual” de un documento pdf. Facilita la navegación por el mismo.

Metadatos: datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas.

OCR (reconocimiento óptico de caracteres): Proceso dirigido a la digitalización de textos, los cuales identifican automáticamente a partir de una imagen símbolos o caracteres que pertenecen a un determinado alfabeto, para luego almacenarlos en forma de datos.

PDF (Portable Document Format): Formato de almacenamiento de documentos digitales independiente de plataformas de software o hardware compuesto de imagen vectorial, mapa de bits y texto. Lanzado como estándar abierto en 2008 (ISO 32000-1).

PDF multipágina: documento pdf formado por la combinación de varios formatos.

PNG (Portable Network Graphics): apareció para solventar las deficiencias del formato GIF y permite almacenar imágenes con una mayor profundidad de color. Está basado en un algoritmo de compresión sin pérdida. Puede llegar a soportar hasta 24 bits.

Picture Manager: software de Microsoft. Permite organizar, editar, compartir y visualizar imágenes.

RAR (Roshal Archive): formato de archivo propietario con un algoritmo de compresión sin pérdida.

Resolución: número de bits que componen la imagen digital.

Software: equipamiento lógico o soporte lógico de un sistema informático que comprende el conjunto de componentes lógicos necesarios que hacen posible la realización de tareas específicas.

TIFF (Tagged Image File Format): archivo estándar para guardar imágenes de alta calidad y muy usado en la impresión de trabajos que utilizan imágenes. De uso común en escáneres. Soporta 48 bits.

XnView: software gratuito para visualizar, convertir y editar imágenes.

ZIP: formato de compresión de archivos informáticos. Es capaz de descomprimir la gran mayoría de archivos comprimidos (zip, rar,...).