UNIVERSIDAD NACIONAL DE LA PLATA

PREBI SEDICI
prebi.unlp.edu.ar sedici.unlp.edu.ar

Manual de Procedimientos de Digitalización PREBI-SEDICI

Esteban C. Fernández, Carlos J. Nusch, Marisa De Giusti, Lorenzo Calamante

Índice

1. Preservación digital: proceso de digitalización

1.1. Acerca de la preservación

El papel es uno de los materiales predilectos en los que circula hoy el conocimiento y, si bien ya está naturalizado, fue un desarrollo tecnológico que superó a sus predecesores, como el papiro y el pergamino. Es probable que pronto se convierta en el predecesor superado de otra tecnología. Sus principales desventajas residen en su propia materialidad (su fragilidad, tamaño y peso, costo). Cada vez más el conocimiento circula en formato digital y uno de los motivos de este cambio paulatino de soporte (además de las ventajas de almacenamiento, transportabilidad, accesibilidad, etc.) es que casi cualquiera puede digitalizar contenidos. En este artículo se darán algunas nociones al respecto.
Cómo digitalizar

1.2. Antecedentes / Normativas

El proceso de digitalización en los repositorios CIC Digital y SEDICI se realiza según estándares reconocidos mundialmente en las áreas de preservación y difusión de archivos digitales.
Se utilizan como guía las directrices definidas en los documentos:

- *"Technical Guidelines for Digitizing Cultural Heritage Materials" generado en 2010 por la Federal Agencies Digitization Guidelines Initiative (FADGI).*
- *"Directrices para proyectos de digitalización de colecciones y fondos de dominio público", IFLA (2002).*
- *"Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files Raster Images", NARA (2004).*
- *“Recomendaciones para la digitalización de los documentos en archivos”. Junta de Castilla y León (2011).*

Versiones

De cada obra digitalizada se obtienen al menos 3 versiones: a) imagen maestra resultante del escaneo, b) documento maestro con imagen derivada de la anterior y el OCR generado sobre ella y c) otra(s) derivada(s) a los fines de uso y distribución en la web.

1.3. El procedimiento de trabajo en SEDICI y en CIC Digital

A partir de las normas, hemos racionalizado en nuestros repositorios un procedimiento de trabajo que presentamos a continuación ordenado en subprocedimientos con sus respectivas instrucciones, de modo que sirvan de modelo para otros repositorios:

En principio, para empezar el proceso de digitalización bastan un escáner plano y una computadora. El trabajo consiste básicamente en tomar un libro o cualquier texto en papel, escanearlo y convertirlo en un archivo digital. Obviamente el proceso es mucho más complejo de lo que suena, sobre todo si se quiere obtener un archivo digital de calidad. Y recordemos que solo vale la pena preservar las buenas digitalizaciones.

El material: A la hora de empezar la digitalización hay que examinar el material. La encuadernación, el estado del papel, el tamaño y la cantidad de hojas son factores que condicionarán el proceso. Por ejemplo, escanear en plano un libro de muchas páginas dificulta una buena captura de los caracteres más cercanos a la unión de las páginas o requiere desarmar la encuadernación. En esos casos es recomendable utilizar un escáner para libros. Si bien el plano funciona mejor con hojas sueltas, también puede procesar, aunque con ciertas dificultades, textos encuadernados.

El escaneo: El proceso de captura del material es el más importante y el más delicado, si no se realiza de manera adecuada lleva, en general, una digitalización defectuosa. Algunos detalles pueden mejorarse con editores de imágenes pero la mayoría de los defectos se mantienen a lo largo del proceso. Un texto borroso o una imagen con poca definición, hojas faltantes, torcidas, dobladas, rotas o manchadas empobrecerán el archivo restándole valor a la conservación.

La edición de la imagen:

El reconocimiento de caracteres: Una vez obtenida la mejor captura posible del texto hay que realizar el reconocimiento óptico de caracteres (OCR, por sus siglas en inglés), para que deje de ser un grupo de imágenes y se convierta en un texto digitalizado. Este proceso consiste en convertir esas letras dibujadas que capturó el escáner en caracteres digitales que puedan ser interpretados por procesadores de texto. Actualmente, existen programas que se encargan de este proceso; es decir que no hace falta introducir el texto con el teclado. Estos programas comparan unos patrones o plantillas de caracteres con las figuras escaneadas y tratan de interpretarlas. Si bien los resultados no son perfectos, con un buen escaneo se pueden obtener excelentes resultados.

Catalogación y archivado: Una vez que se ha conseguido un texto digital con buenas imágenes y un buen reconocimiento hay que decidir en qué formato se va a guardar para evitar, tanto como sea posible, la obsolescencia. El formato es algo así como el soporte material del texto digital, por eso mientras más estable sea mejor preservado estará. Además, también es importante catalogarlo de tal manera que sea fácilmente ubicable y reconocible para que no se pierda en la proliferación constante de archivos digitalizados. También es conveniente tomar la precaución de tener copias de seguridad de los archivos para evitar pérdidas por deterioro de hardware.
Cada uno de estos procesos implica una gran cantidad de complejidades técnicas (que serán abordadas en próximas publicaciones). En este artículo quisimos dar un breve panorama de las etapas del proceso de preservación digital.

1.3.1. Primera parte - Recepción, análisis y evaluación del material a digitalizar

Todas las obras antes de ingresar al flujo de trabajo son evaluadas teniendo en cuenta estos criterios:

- Estado general de conservación
- Dimensiones
- Formatos
- Tipos de encuadernación
- Importancia histórica, educativa, institucional

El tiempo que se destina al proceso de digitalización depende directamente de los primeros cuatro puntos antes mencionados. Es por esto que, teniendo en cuenta su importancia, ingresan o no al flujo de trabajo.

1.3.2. Segunda Parte - Elección de metodología de escaneo

El material se revisa y caracteriza según su estado, teniendo en cuenta la encuadernación, el estado del papel, el tamaño y la cantidad de hojas. De estos factores depende la elección del escáner más indicado para asegurar la correcta manipulación del ejemplar físico y obtener las imágenes maestras de la mayor calidad y fidelidad posible.

1.3.2.1 Los escáneres de SEDICI y CIC Digital

En SEDICI y CIC Digital contamos con cinco escáneres:

1.- **Un escáner automático Fujitsu FI 7160**

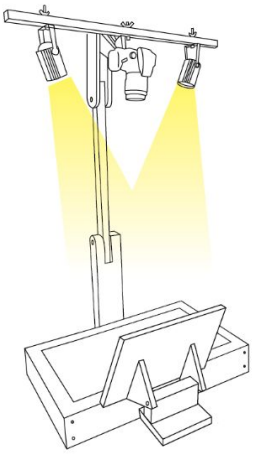
2.- **Un escáner HP 7500** que dispone de dos opciones de escaneo: una cama plana y un alimentador automático.

Escáner automático

3.- **Un escáner Contex IQ Quatro** que permite escanear tamaños de hasta 44 pulgadas (111.75 cm) por lo que es un equipo ideal para el caso de planos.

Escáner de gran formato

4.- **Un escáner cenital** que se construyó a partir de la modificación de uno de los modelos disponibles en https://www.diybookscanner.org/. Cuenta con una cámara reflex Nikon D5600 con un lente del tipo zoom 18-55mm y un formato de sensor DX de 24,2 Megapíxeles que permite obtener documentos escaneados a 400dpi y dos lámparas LED dicroicas de luz cálida, cuya temperatura no daña el material y que consumen hasta un 90% menos, no generan calor considerable y tienen un Índice de reproducción cromática (CRI)[1] alto (FADGI, 2016: 17).



5.- **Un escáner DAL Archivista 2014:** Este escáner fue también construido a partir de los modelos disponibles en https://www.diybookscanner.org/. Cuenta con dos cámaras Nikon y dos lámparas que tienen un CRI alto. El libro se apoya sobre una cama móvil que permite abrirlo a 90º y que, accionada por una palanca, hace tope con un cristal que permite "alisar" la página.

[1] El índice de reproducción cromática (IRC) es la medida de la capacidad de una fuente de luz para mostrar los colores de un objeto de manera "real". Se trata de una escala del 1 al 100, tomando como valor de referencia y más alto la iluminación natural.

1.3.2.2. La selección del escáner

1. **“Material a digitalizar/Puede ser desarmado/Tamaño máximo hoja oficio/Fujitsu FI 7160 ó HP 7500 ó Escáner cenital”)**

Si el material se encuentra en buen estado, utilizamos los escáneres Fujitsu o HP. Las hojas separadas no deben tener bordes sobresalientes, ni dobleces que puedan generar atascos en el alimentador automático.
Si se elige el escáner Fujitsu FI 7160 el software a utilizar es el **PaperStream Capture** o el **ABBY Fine Reader** y en el caso del HP 7500 el **“Software de escaneo de documentos inteligente HP”** contemplando la configuración de captura con estos parámetros:

Imágenes TIFF sin pérdida de 400 pp. para las páginas a color y escala de grises y 600 pp. para las páginas en blanco y negro.

En cambio si el papel se encuentra friable, utilizamos el escáner cenital, con el que obtendremos imágenes JPG sin pérdida, de 3000 píxeles de ancho por 4000 de alto.

Si es necesario desarmar, luego habrá que encolar, anillar o abrochar según el caso.

2. **“Material a digitalizar/Puede ser desarmado/Grandes tamaños/Contex IQ Quattro** (para el caso de planos, mapas etc)

El escáner Contex IQ Quattro permite escanear tamaños de hasta 44 pulgadas (111.75 cm) por lo que es un equipo ideal para el caso de planos. El software que controla el equipo es el **NextImage**

3. **“Material a digitalizar/No puede ser desarmado/Tamaño mayor a hoja A4/Escáner cenital**

El material a procesar en este caso puede estar o no en buen estado, la principal característica es el gran tamaño y la imposibilidad de que sea desarmado. El software a utilizar es el **digiCamControl.**

4. **“Material a digitalizar/No puede ser desarmado/Tamaño máximo hoja A4/Buen estado de conservación/Escáner DAL**

Seleccionaremos este escáner si el material es sensible y no puede tener una apertura de 180 grados. El tamaño máximo admitido por el escáner DAL es de 33X20cm, el software de escaneo a utilizar es el digiCamControl.

5. **“Material a digitalizar/No puede ser desarmado/Tamaño máximo hoja A4/Buen estado de conservación/HP 7500 plano**

Este es el caso más utilizado cuando el material no sufre deterioro si es abierto 180 grados. Suelen ser el caso de revistas, boletines etc.

5. **“Material a digitalizar/No puede ser desarmado/Tamaño máximo hoja A4/Mal estado de conservación/Escáner cenital**

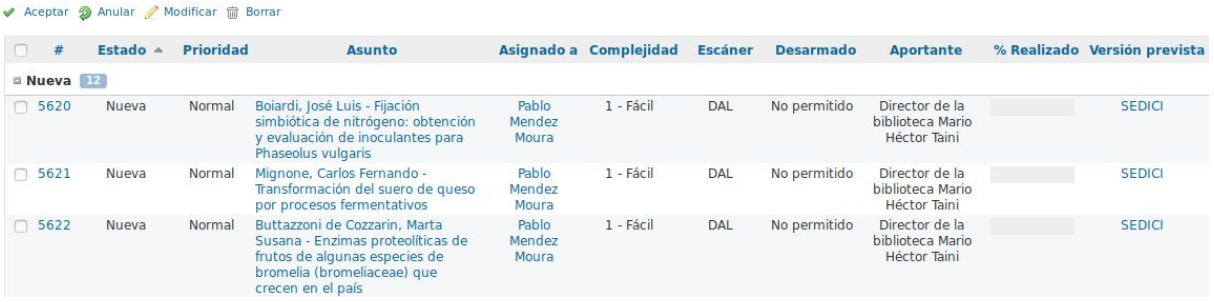
En este caso, por el estado del material, no es posible utilizar el escáner Dal ya que podría dañar las hojas. El software a utilizar para la captura es el **digiCamControl**.

El material encuadernado que no puede desarmarse es escaneado a través del escáner de libros DaL. Estas imágenes son JPG sin pérdida, de 3000 píxeles de ancho por 4000 de alto. Luego de ser capturadas con las cámaras deben combinarse pares e impares, renombrarse y rotarse. En todo momento se intenta mejorar la captura de los materiales que por distintos motivos pueden llegar a estar deteriorados. Este proceso de preparación es fundamental para la obtención de la mejor imagen maestra posible.

1.3.3. Tercera parte - Carga de materiales en el sistema de gestión (Redmine)

Para la gestión de los distintos estados y avances de los materiales ingresados para digitalizar se utiliza Redmine. Este sistema permite realizar un seguimiento completo de todas las etapas que atraviesan dichos materiales.
Luego de tener en claro todas las particularidades de cada caso se:

- asigna el estado de conservación del material.
- selecciona el escáner apropiado de acuerdo al formato
- asigna una persona responsable
- determina la complejidad (Fácil, medio o difícil)
- agregan todos los datos propios del material (Autor y título, aportante y lugar de origen)
- Para los procedimientos en los que estructuramos este manual, las tres primeras partes corresponden al estado “Nueva”; la cuarta a “En captura”, la quinta a “En edición”, la sexta a “En postproceso” y la séptima a “para subir”. Una vez subido a SEDICI, pasa a “para devolver” y luego a “Resuelta”.
  Si por algún motivo nos vemos obligados a detener el trabajo, dejamos constancia mediante la opción “comentarios” y si no podemos resolverlo, pasa a “Rechazada”.


Aceptar Anular Modificar Borrar

#	Estado	Prioridad	Asunto	Asignado a	Complejidad	Escáner	Desarmado	Aportante	% Realizado	Versión prevista
Nueva 12										
5620	Nueva	Normal	Boiardi, José Luis - Fijación simbiótica de nitrógeno: obtención y evaluación de inoculantes para Phaseolus vulgaris	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI
5621	Nueva	Normal	Mignone, Carlos Fernando - Transformación del suero de queso por procesos fermentativos	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI
5622	Nueva	Normal	Buttazzoni de Cozzarin, Marta Susana - Enzimas proteolíticas de frutos de algunas especies de bromelia (bromeliaceae) que crecen en el país	Pablo Mendez Moura	1 - Fácil	DAL	No permitido	Director de la biblioteca Mario Héctor Taini		SEDICI

A medida que las obras van pasando por distintas etapas, también se verá reflejado en el sistema hasta que el proceso finaliza.

1.3.4. Cuarta parte - Captura de imágenes

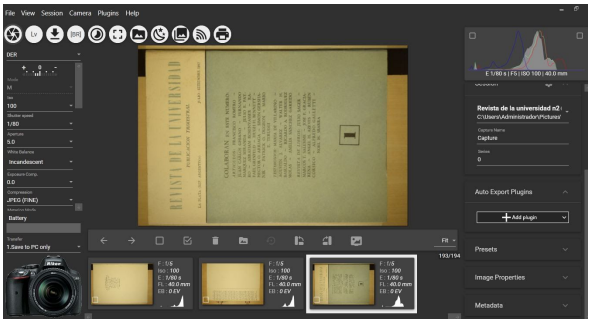
De acuerdo al escáner que haya sido seleccionado se utilizará un determinado software de captura:
1. Si el escáner seleccionado es el Fujitsu FI 7160 (de alimentación automática) se utiliza tanto **PaperStream** como **ABBY FineReader**
2.El escáner Hp 7500 dispone de dos opciones de escaneo una cama plana y un alimentador automático que se controlan mediante el **Software de escaneo inteligente HP.**

3. El escáner de gran formato Contex IQ Quattro 4400 se gestiona con el software **Nextimage**
4. Para controlar, configurar y realizar la previsualización en el momento de la captura con las cámaras réflex de los escáneres DAL Archivista 2014 y el cenital, se utiliza el software **DigiCamControl**.

1.3.4.5. Captura con "digiCamControl"[2]

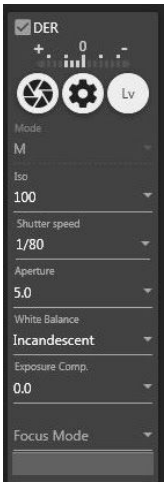
Este software permite la configuración y control completo de las cámaras que se utilizan tanto en el escáner DaL como en el cenital.

Visualización instantánea



Las imágenes obtenidas se visualizan inmediatamente en pantalla completa, permitiendo de esta manera visualizar rápidamente problemas de encuadre, foco, iluminación etc.

Control avanzado de captura



Permite configurar velocidad de obturación, apertura de diafragma,ISO, balance de blancos.

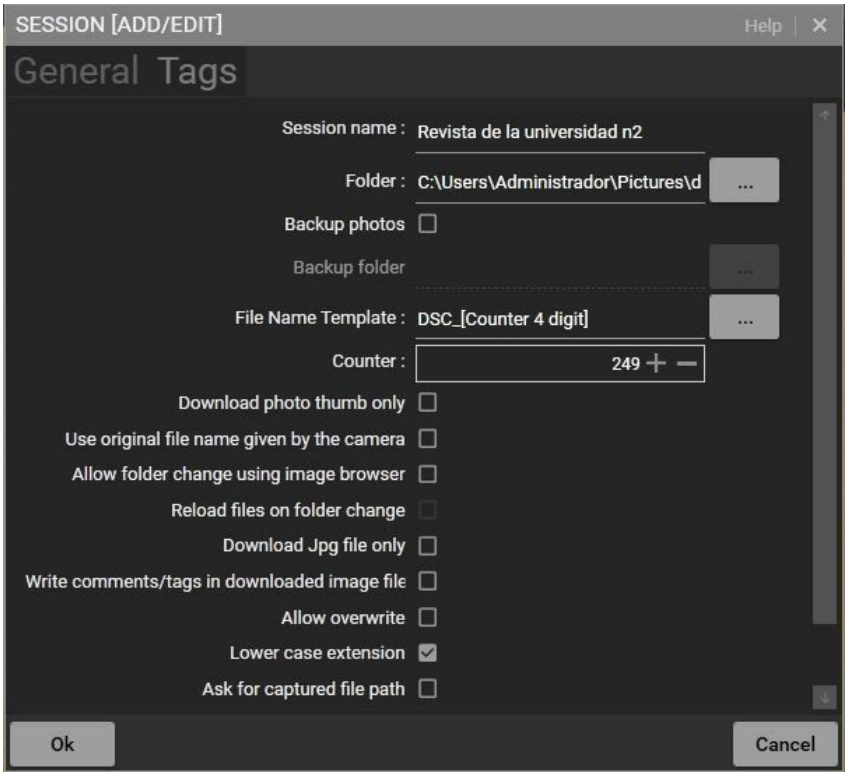
Soporte de cámara múltiple

Puede controlar varias cámaras conectadas al mismo tiempo, activando la captura de fotos en paralelo, o una por una.

[2] Este instructivo puede ampliarse con la lectura del manual completo de este software en: http://digicamcontrol.com/doc/userguide

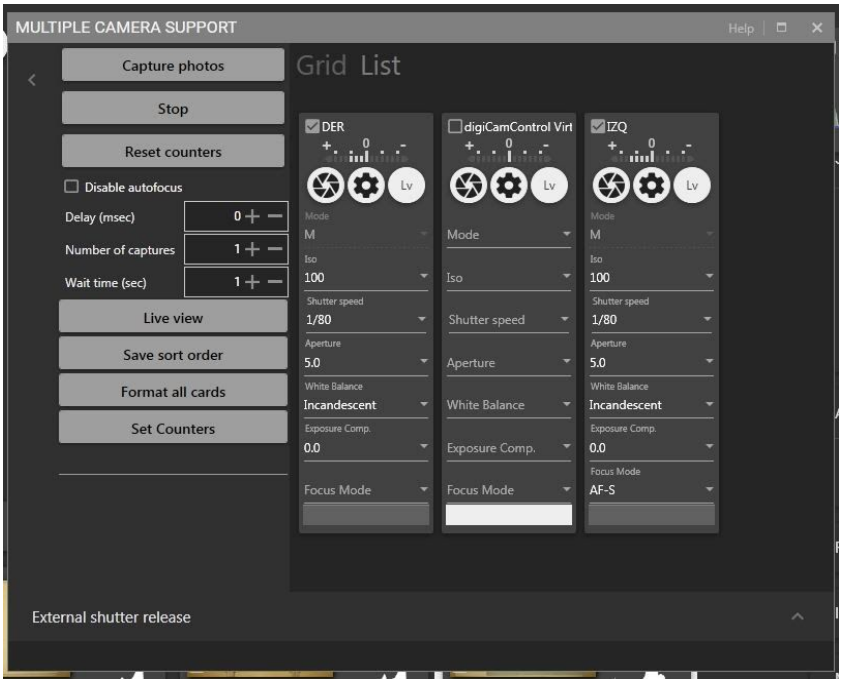
Configuración Inicial:

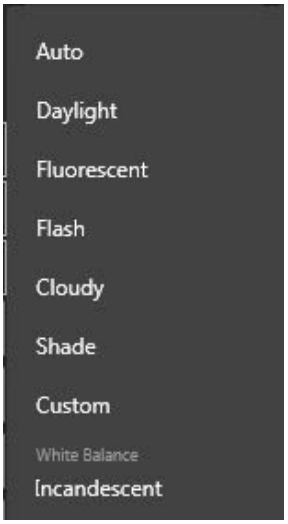
1. **Crear una nueva sesió**n, desde aquí podemos configurar el nombre del proyecto y donde se guardaran las imágenes capturadas.



2. **Configuración de las cámaras:**

La opción Multiple Camera Support  permite acceder a los ajustes de las cámaras conectadas. Los ajustes necesarios antes de comenzar el proceso de digitalización son los siguientes:

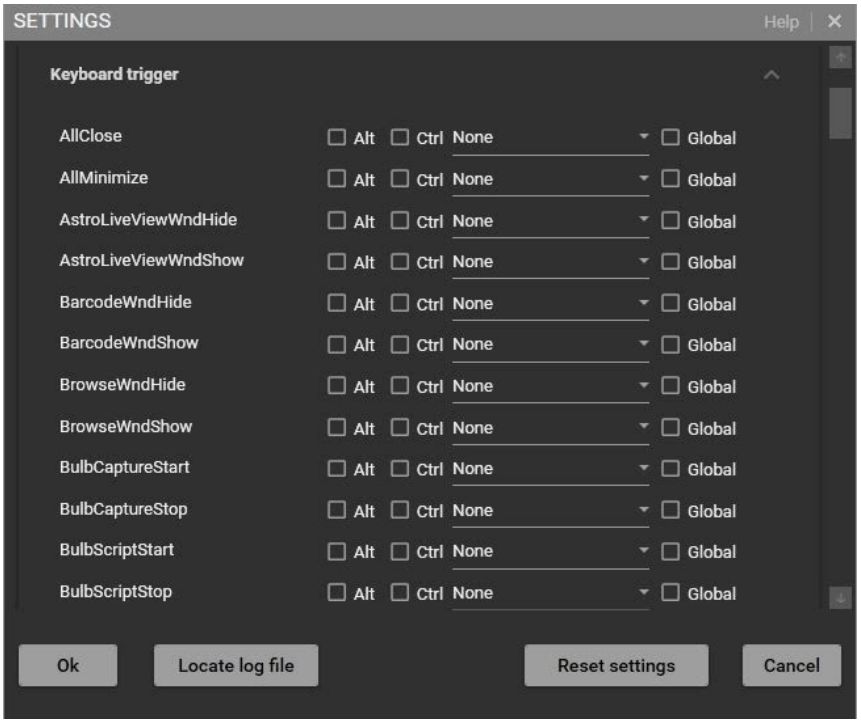




- **ISO**: es la capacidad que tiene el sensor de la cámara para captar luz. A valores bajos de la sensibilidad **ISO**, el sensor es menos sensible a la luz. Es recomendable tratar de utilizar el valor más bajo ya que a valores altos hay más interferencia/ruido en las imágenes
- **Shutter speed** (Velocidad de obturación): Es el tiempo en que queda expuesto el sensor en la captura. Cuanto más tiempo quede expuesto más luminosa será la imagen pero si los valores son más bajos de 1/40 puede aparecer la imagen movida
- **Aperture** (Apertura del diafragma): La apertura del diafragma incide directamente en la cantidad de luz que ingresa. Cuanto más cerca de 0 más luz ingresa
- **White balance** (Balance de blancos) Es un ajuste que se utiliza para representar los colores con mayor fidelidad. Los distintos artefactos de iluminación tienen distinta temperatura de color (tonalidad de la luz) y en el momento de la captura se selecciona la opción que representa mejor a la tonalidad real.

3. **Configuración de teclado** para: captura, visualización en vivo, acceso a configuraciones

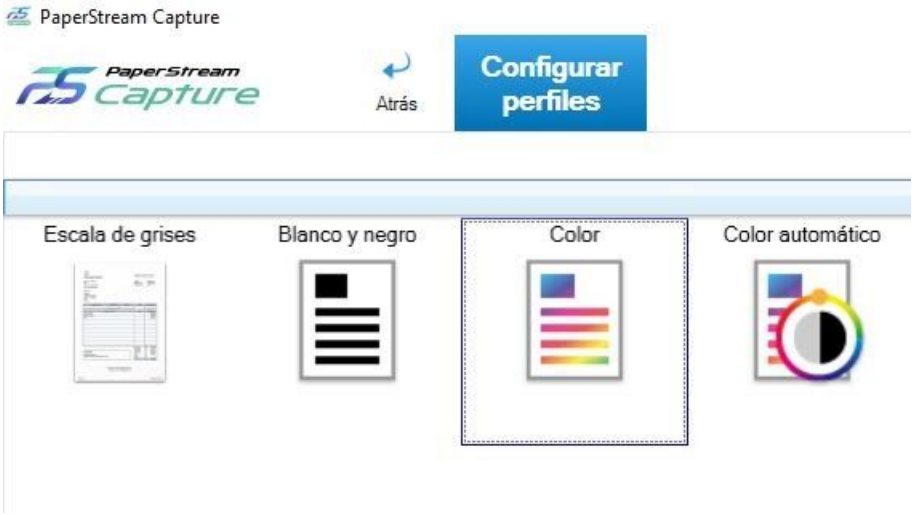
Acceso desde File - Settings - Keyboard trigger



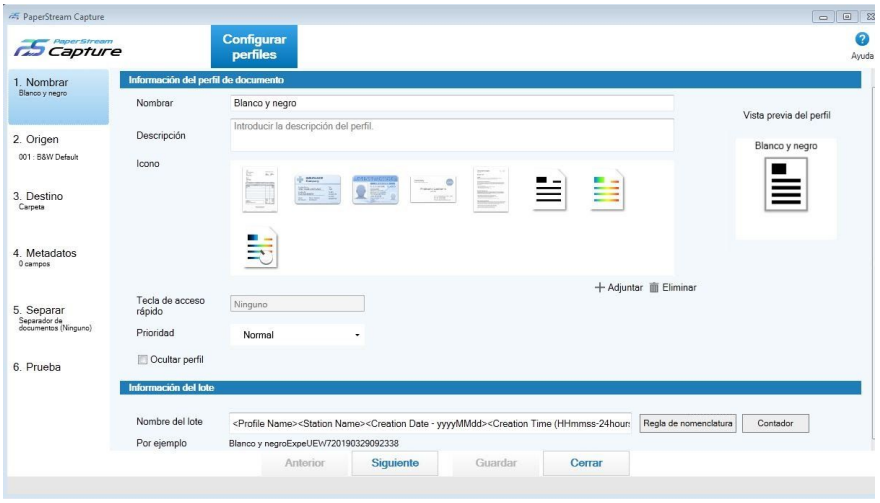
En este ejemplo se observa cómo asignar el número 0 del teclado numérico para iniciar la captura de las cámaras.

1.3.4.1. Captura con “Paperstream”

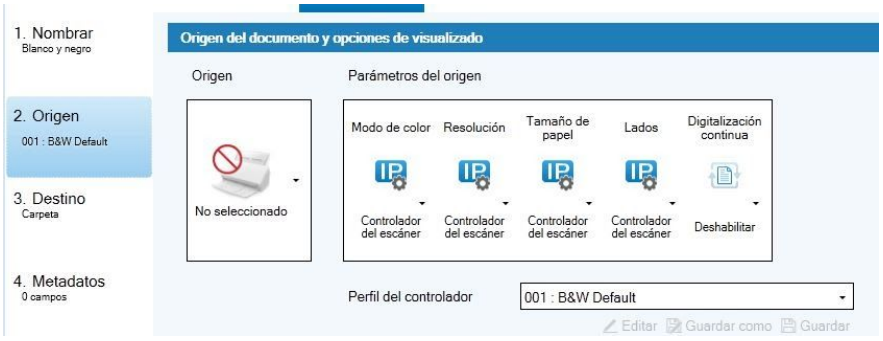
Este es el software utilizado para la captura con el escaner Fujitsu FI 7160. Desde la pantalla inicial se accede a los perfiles pre definidos con la configuración adecuada a cada tipo de material.



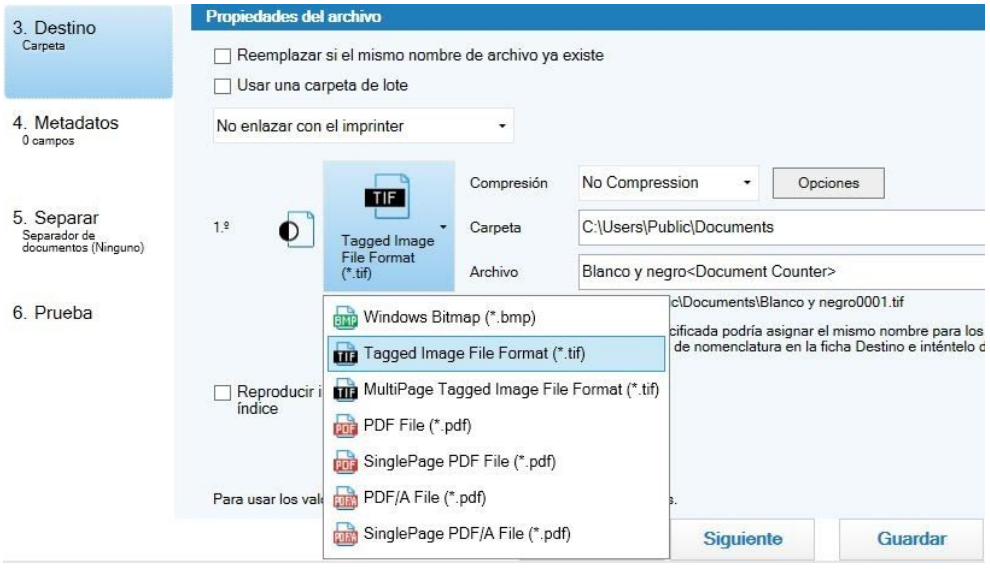
Se configuran los perfiles de digitalización teniendo para que se adapten a cada necesidad. En principio son tres “Blanco y negro”, “Escala de grises” y “Color”



Desde la pestaña “Origen del documento y opciones de visualizado” se configura la resolución en el caso de blanco y negro es 600 DPI en Escala de grises es 400 DPI y por último en Color es de 400 DPI.



En la pestaña “Destino” se configura el formato en que será guardado el documento. Formato Tif sin compresión



En la pestaña “Separar” se configura la detección de los elementos que van a determinar cuando termina un documento, para de esta manera, poder ser procesados por lote.

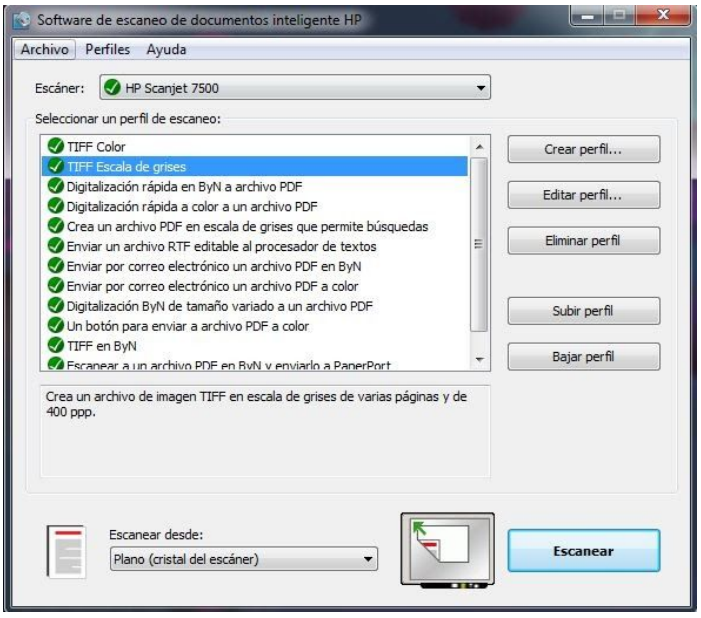
Una herramienta muy útil que dispone PaperStream es el “Administrador de lotes” este permite digitalizar muchos documentos distintos separados por un identificador que puede ser:

- Página en blanco
- Contador de páginas
- Zone OCR: detecta una valor dentro de una zona del documento pre fijada
- Patch code: permite detectar este tipo de marcas colocadas previamente en sobre los documentos.
- Códigos de barras
- Pulsar CTRL

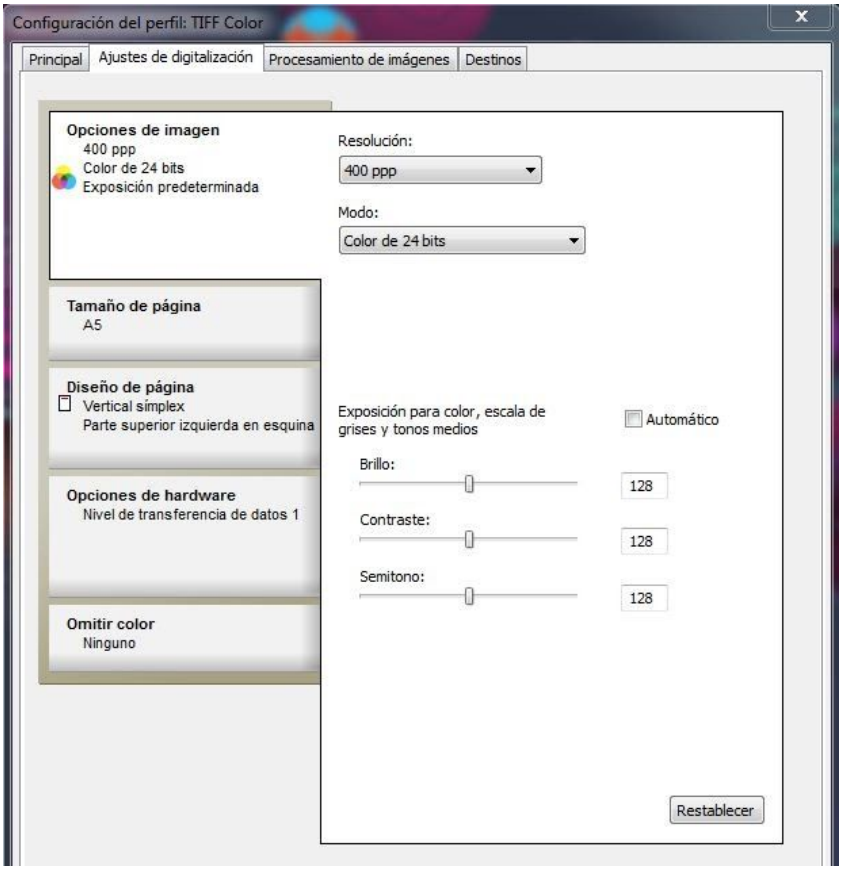
1.3.4.2. Captura con ABBYY FineReader

1.3.4.3. Captura con “Software de escaneo de documentos inteligente” de HP (Versión 3.7.1)

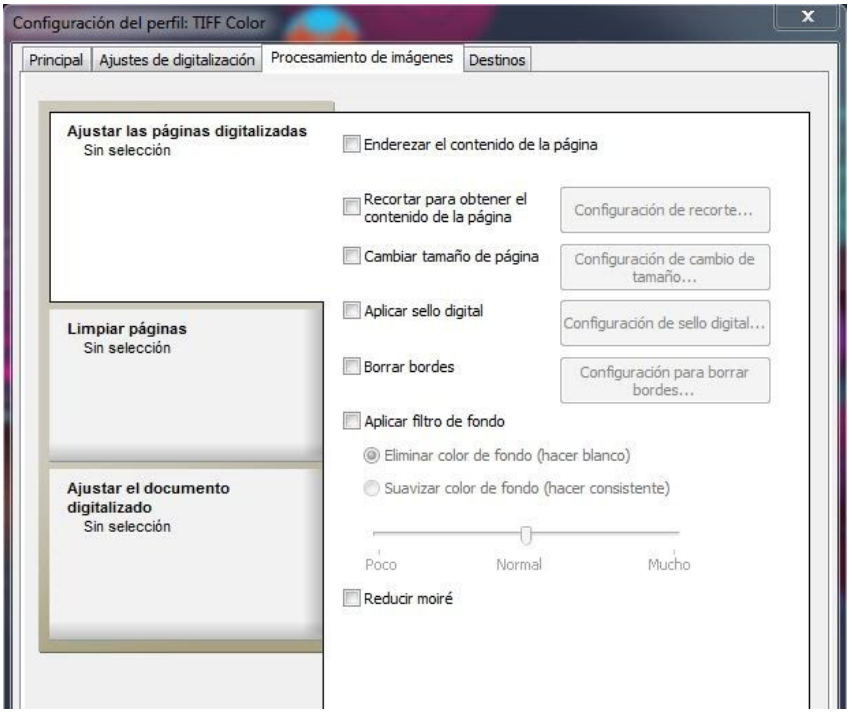
La pantalla principal lista los perfiles pre definidos que pueden ser utilizados y una lista de botones para crear, editar, eliminarlos.



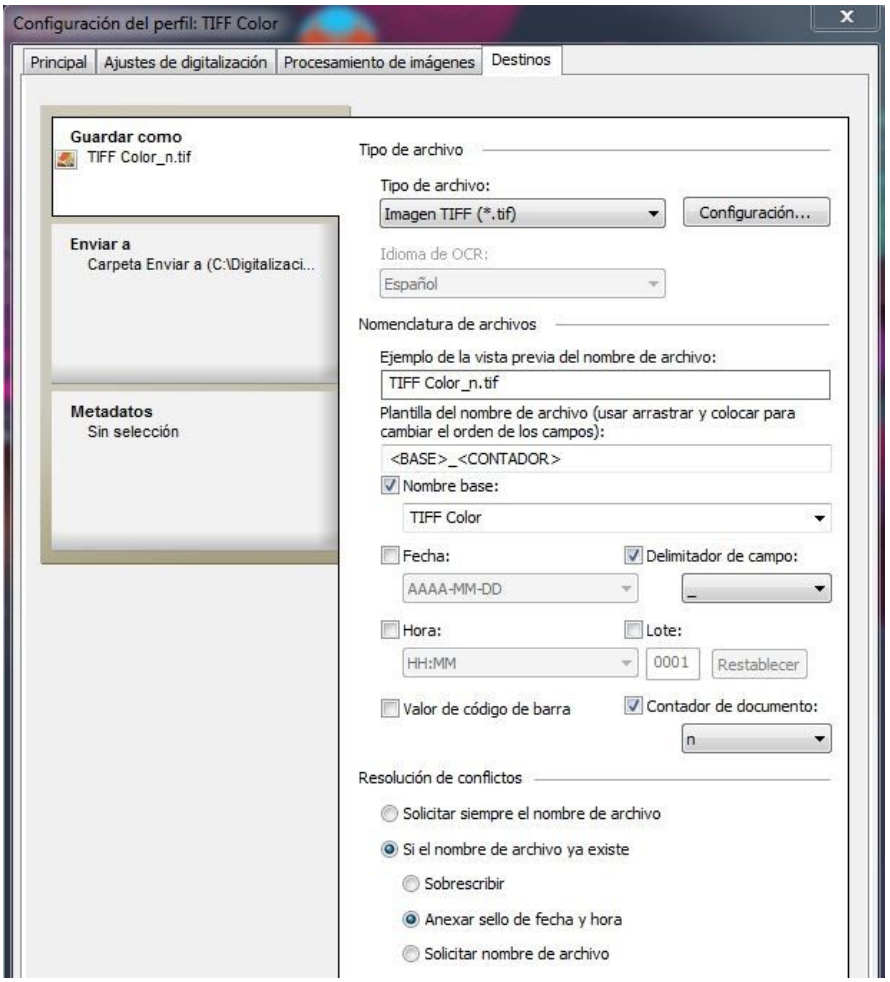
En las opciones para crear y editar perfiles se encuentran los ajustes para la digitalización. Para la captura en color y escala de grises se configura en 400 DPI y en blanco y negro 600 DPI.



La pestaña procesamiento de imágenes tiene opciones para la mejora de la captura. Este proceso se aplica a cada página a medida que va realizando la digitalización y no puede ser revertido.



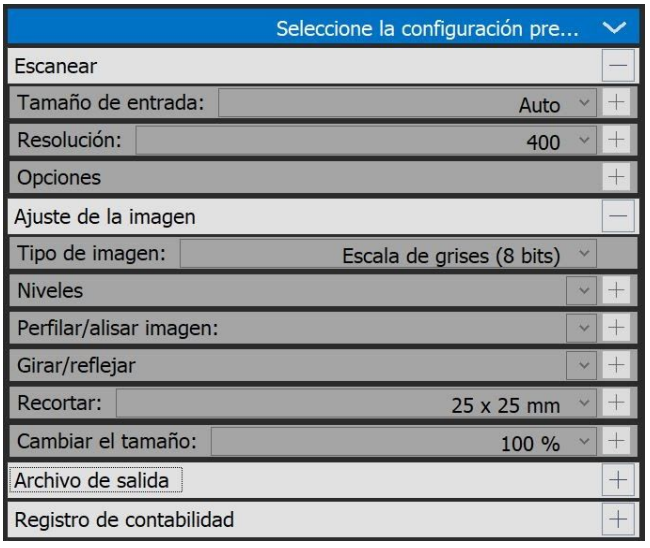
La pestaña “Destinos” permite configurar el formato de salida del documento y ajustar el tipo de compresión. En todos los casos el tipo de compresión a utilizar es “sin perdida”



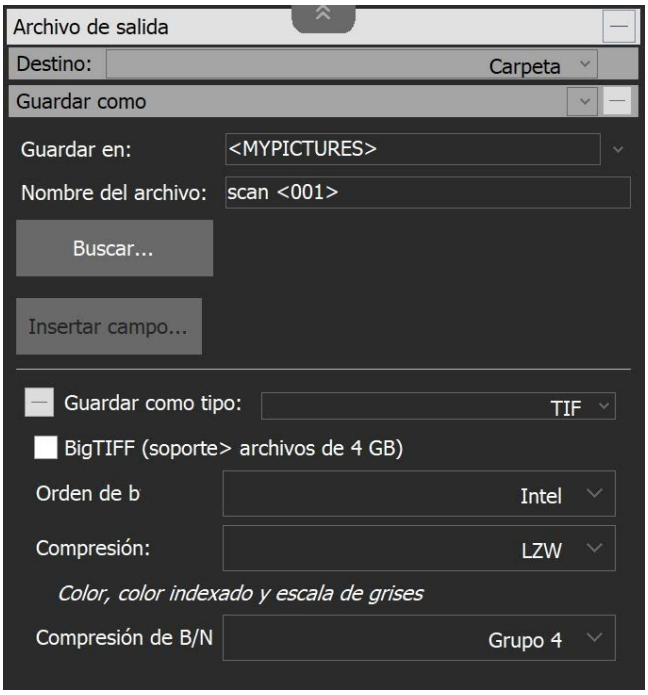
1.3.4.4. Captura con “NextImage”

Este software controla el escáner Contex IQ Quattro

Sobre la derecha se encuentra el panel de configuración donde se establecen los parametros de escaneo.



Dentro de este panel casi al final de la lista se encuentran las opciones de guardado



En la parte superior central se encuentran los Botones Escanear - Copiar - Archivo
El botón Escanear ejecuta el escaneo con las opciones pre configuradas
El botón Copiar ejecuta el escaneo y luego envía a imprimir
El botón Archivo

1.3.5. Quinta parte – Edición de imágenes

Una vez finalizada la captura se realizan procesos de edición y mejora de las imágenes obtenidas.

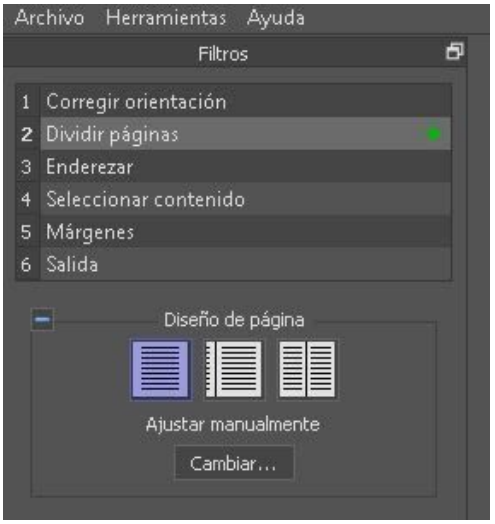
1.3.5.1. Edición con ScanTailor Advanced

Scantailor es una herramienta gratuita de código abierto que permite corregir o modificar las imágenes capturadas. Soporta los siguientes formatos de entrada: *.tif, *.tiff, *.png, *.jpg, *.jpeg y genera archivos con formato tiff de salida (hasta dos por cada página).
Algunas de las opciones más importantes son:

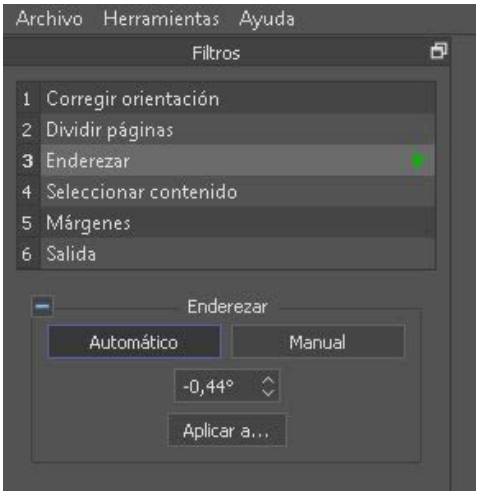
1. **Corregir PPP**
   Esta es una función necesaria para corregir los por puntos por pulgadas o DPI de las imágenes maestras obtenidas a través del escáner DaL o cenital (para las capturas de escáneres planos o automáticos este proceso no es necesario).

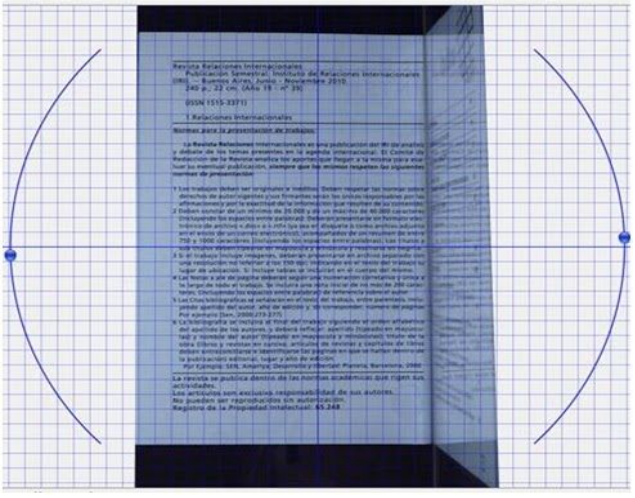
2. **Corregir orientación**
   Puede ser necesario rotar todas o alguna de las páginas capturadas.

3. **Dividir páginas**
   Permite dividir las capturas en las que ven hasta dos páginas.



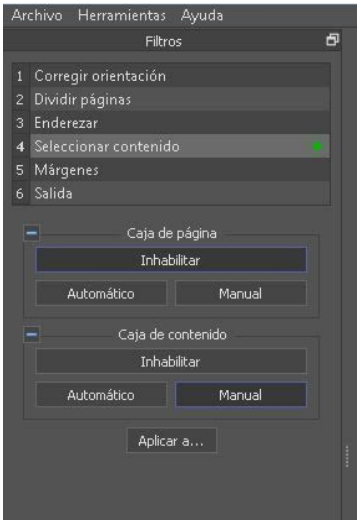
4. **Enderezar**
   Esta herramienta permite encuadrar el texto y la imagen a 90º; puede configurarse tanto automáticamente como manualmente.





5. **Seleccionar contenido**

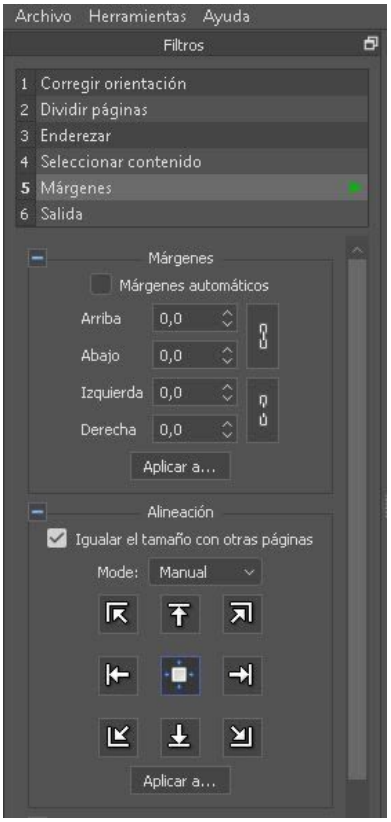
Esta etapa determina la región con contenido "útil" del documento. El límite exterior de los márgenes afecta el tamaño del archivo de salida y define el tamaño que tendrán las páginas (sin contar los márgenes. Las imágenes finales no muestran la línea de plegado u otros restos de los bordes.



6. Definir márgenes

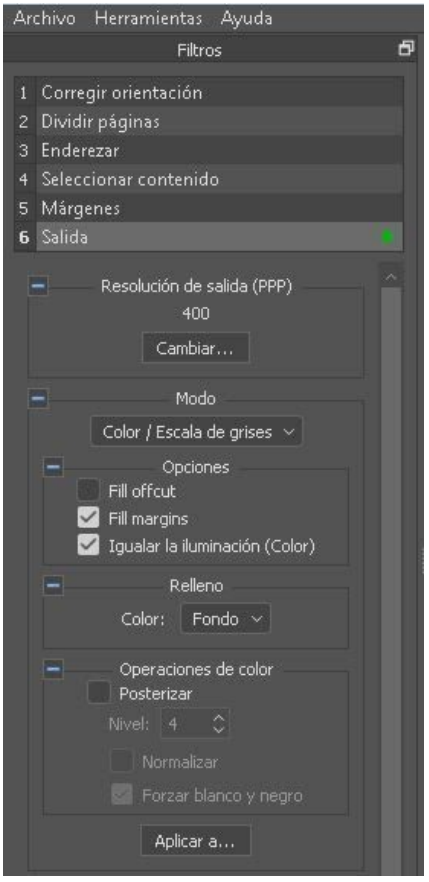
En esta etapa, se ajustan los márgenes agregados al área de contenido.

La opción “Igualar el tamaño con otras páginas” asegura que las páginas tengan el mismo tamaño mejorando así la visualización general del documento.



7. Salida

En esta etapa, los archivos de salida se crean a partir de las modificaciones realizadas a lo largo de todas las etapas y se guardan en el disco. A diferencia de las otras etapas, "Salida" está disponible solo después de que todas la páginas pasen las etapas de "Seleccionar contenido" y "Márgenes" ya que es necesario haber determinado el tamaño de las páginas. En esta etapa pueden controlarse las siguientes variables:



Resolución de salida (PPP) se define la resolución de los archivos (normalmente se define la imagen a 400 ppp o dpi).

Modo: tres modos de tratamiento de color están disponibles:

Blanco y negro usado para los casos que solo hay texto ya que como imagenes no se las representa bien.

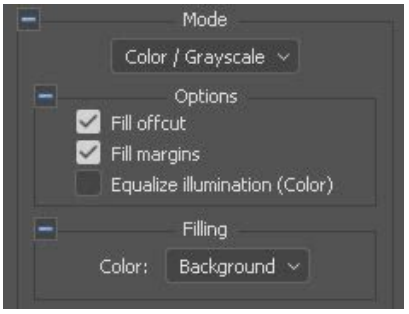
Color / Escala de grises utilizado para páginas con textos que presenten tipografías difusas, textos deteriorados y páginas que contengan imágenes.

Mezclado: esta opción mezcla el modo blanco y negro con el Color / Escala de grises detectando manual o automáticamente las páginas que contengan imágenes, dejando en color estas y los textos en blanco y negro

La opción Fill offcut es utilizada para rellenar los márgenes que se generan cuando se dividen las páginas.

La Opción Fill margins se utiliza para rellenar los márgenes que son generados en el proceso “Márgenes”.

El color del relleno puede ser blanco o una muestra del color de fondo de la página.



Cada modo tiene opciones distintas para el tratamiento de las imágenes. Las opciones “Equalize illumination (B&W)” y “Equalize illumination (color)” realizan un balance entre los colores más intensos y el blanco y entre blanco y negro.

La opción “Threshold” corresponde al **método de valor umbral**, por medio del cual la imagen pasa a blanco y negro y que explicaremos muy someramente a continuación: el programa convierte la imagen a escala de grises, detecta los valores de gris más alto (ésto es, más cercanos al negro) y los más bajos (más cercanos al blanco) y determina la media a partir de la cual los valores más altos se convertirán a negro y los más bajos a blanco, pudiendo ser modificada manualmente.
Los métodos Sauvola y Wolf (anteriormente explicamos el Otsu) operan del mismo modo, pero segmentando la imagen por regiones, de modo tal que, con con distintos valores de mayor gris y menor blanco por región no se pierda información valiosa, cosa que sí pasaría si se procesara la totalidad de la imagen.

La herramienta “despeckling” reconoce en los modos B&W y Mixed los puntos que podrían estar produciendo ruido en la imagen, tomando como parámetros el tamaño del punto y su cercanía con otros objetos de mayor tamaño (de modo tal que no se pierdan, por ejemplo, los signos de puntuación).

La herramienta “Fill Zones” permite trazar un polígono o dibujar una forma irregular que cubrirá la imagen, ya sea con blanco o con negro, como con cualquier otro color que con el cuentagotas de la opción “pick color” se seleccione de la imagen.
https://github.com/scantailor/scantailor/wiki/C.-Output-Tabs:-Despeckling-&-Fill-zones
La herramienta “dewarping” permite corregir -tanto manual como automáticamente- la distorsión trapezoidal y la curva producida por la apertura del libro.
https://github.com/scantailor/scantailor/wiki/B.-Output-Tabs:-Dewarping
Mediante la herramienta “splitting”, puede recuperarse el color de algunos trazos en el modo B&W y posterizar una imagen en modo Color/Grayscale; esto es, someramente, la eliminación de tonos de los colores.

1.3.5.2. Edición de imagen con ABBY FineReader

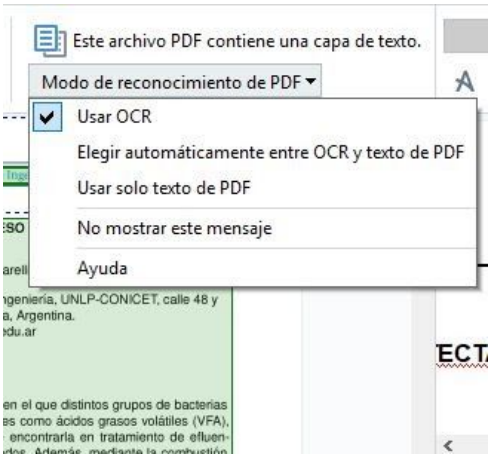
ABBY FineReader permite también editar las capturas, proponiendo el software un prepocesamiento predeterminado o bien poniendo a disposición del usuario una serie de herramientas que permiten dividir la imagen, rotarla, desinclinarla, corregir la distorsión trapezoidal y pasar a blanco y negro. De este modo pueden también resolverse problemas de visualización que dificulten el OCR durante el posproceso.

1.3.6. Sexta Parte - Reconocimiento de caracteres (OCR)

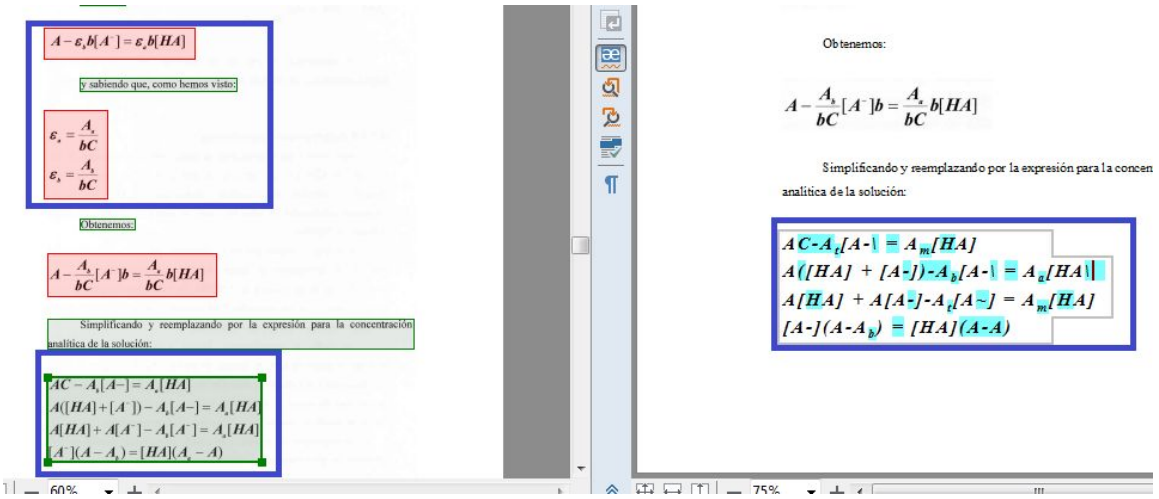
Para este proceso se utiliza Abbyy FineReader, este software permite realizar un reconocimiento óptico de caracteres, posee un editor de texto donde se corrige manualmente las palabras que contienen errores y por último los archivos se guardan en formato PDFA

Al cargar los archivos, el software analiza el contenido y realiza el reconocimiento del texto. Para esto detecta las imágenes, las tablas y los textos creando un recuadro sobre cada selección. Esta tarea también puede realizarse de manera manual apretando el botón Analizar página y Reconocer página.

Algunos documentos generados con distintas herramientas pueden tener errores en la capa que contiene el texto. En el momento del reconocimiento es posible seleccionar si se realiza un OCR nuevamente del documento, si se toma el ocr que trae el documento original (si es que lo tiene) o si el software lo hace de manera automática.



Las fórmulas matemáticas y estadísticas son seleccionadas como imágenes ya que el OCR no puede interpretarlas correctamente la mayor parte de las veces.



Luego se realiza el reconocimiento ópticos de caracteres (OCR) y se lo revisa y corrige manualmente, poniendo especial cuidado en los índices y las portadas, donde se encuentra información importante, como es el título completo, los autores -así como también otras personas involucradas (por ejemplo, directores de tesis)- y la fecha de publicación.
El OCR permite la búsqueda, el copiado y la transformación del texto por el usuario final así como la extracción e indexación por medio de crawlers.

1.3.7. Séptima parte - Guardado de archivos y reglas de nomenclatura

Tres tipos de archivos son guardados

- Archivos maestros sin edición de imagen y guardados en formato Tiff con compresión sin pérdida
- Archivo Pdf/A las imágenes capturadas son editadas para solucionar problemas de inclinación, borrar manchas, mejorar la legibilidad, normalizar tamaño de páginas y con reconocimientos de caracteres (OCR)
- Archivo Pdf/A con compresión para divulgación además de la edición y OCR se comprime para generar un archivo de fácil descarga y visualización

El proceso final incluye el guardado del archivo de Abbyy Fine Reader y la generación de dos archivos PDF/A1, uno sin compresión con el fin de conservación y el otro comprimido para fines de uso y distribución web. En los casos en que es necesario (algunos libros y colecciones) se divide el archivo PDF para uso y distribución web en artículos individuales. Según las posibilidades del software de transformación se genera PDF/A1 (ISO 19005) o PDF/A2 (ISO 320001). Siempre que sea posible, se generan formatos PDF/A1a y PDF/A2a dado que garantizan la accesibilidad, correcta visualización y permiten preservar las características estructurales del documento y el OCR generado. Cuando esto no es posible se genera PDF/A1b o PDF/A2b.

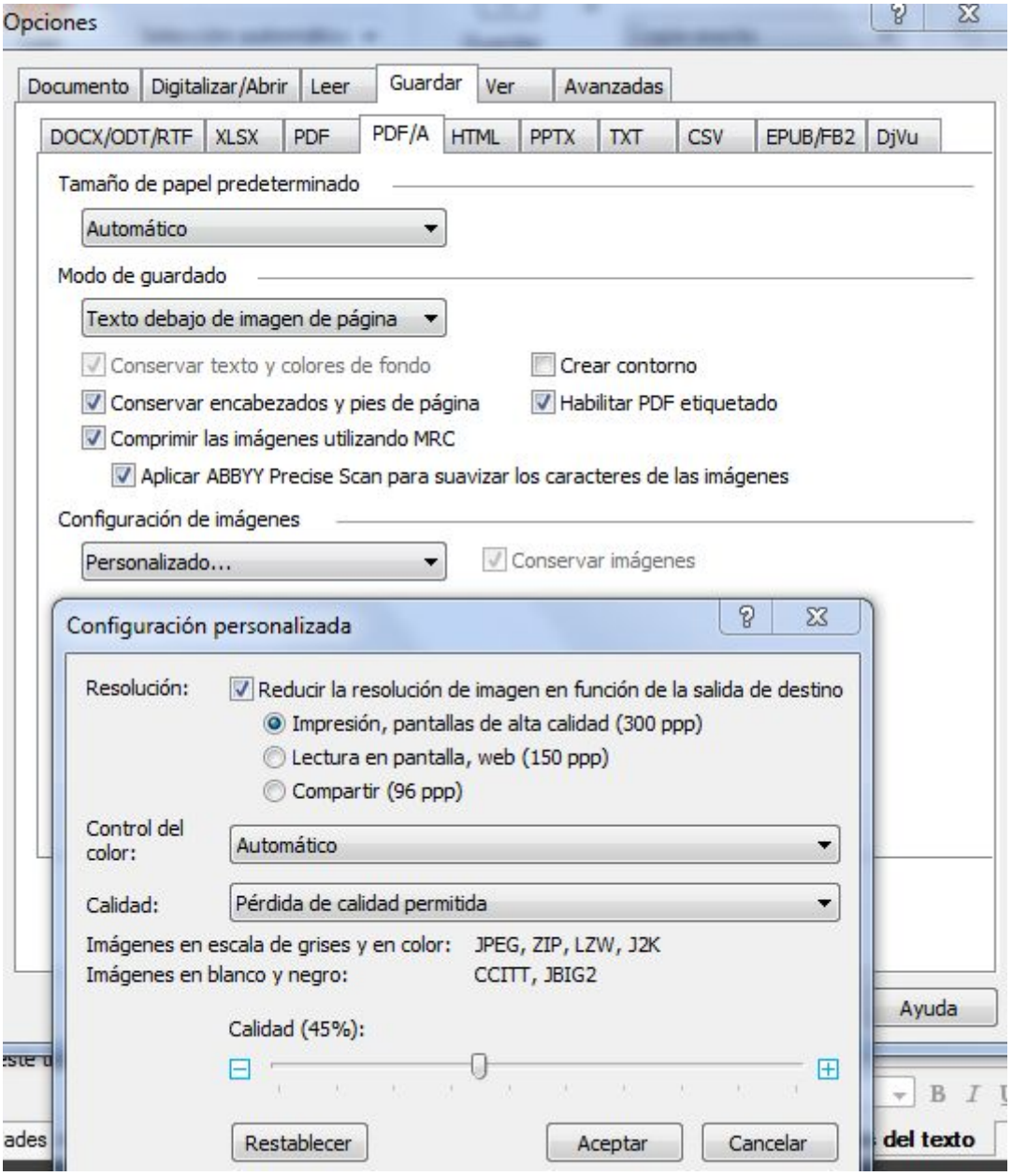
¿Qué es el PDF/A?
El PDF/A es un formato de archivo para el guardado a largo plazo de documentos electrónicos. Debe seguir ciertas reglas incluyendo la conformidad con las directrices en

cuanto a la gestión de color, el uso de fuentes integradas a la hora de la visualización, o la posibilidad de realizar anotaciones por parte del usuario.

El documento muestra la configuración de los perfiles de pdf/A para guardar el documento una vez terminada su edición. Luego, optimizar desde acrobat.

Documento comprimido

Este será, efectivamente, el que se suba completo. Debe ser pdf/A y optimizarse con Acrobat.
La configuración de la compresión en ABBY FineReader debe ser de la siguiente manera:

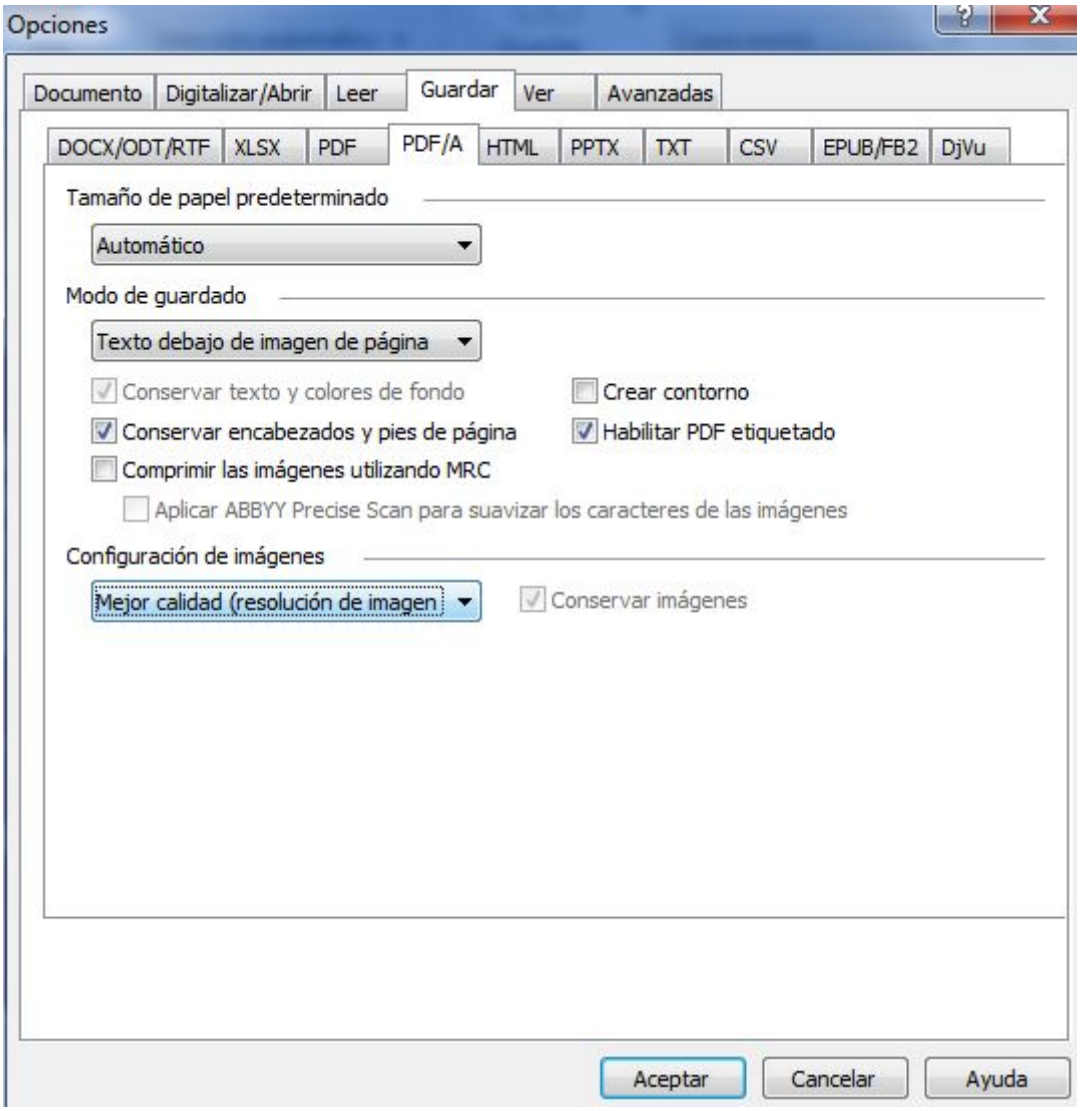


Imágenes en personalizado - resolución 300 ppp - pérdida permitida.

Documento calidad original

Una vez que se terminan todas las correcciones de edición/ocr, se genera el original sin pérdida.

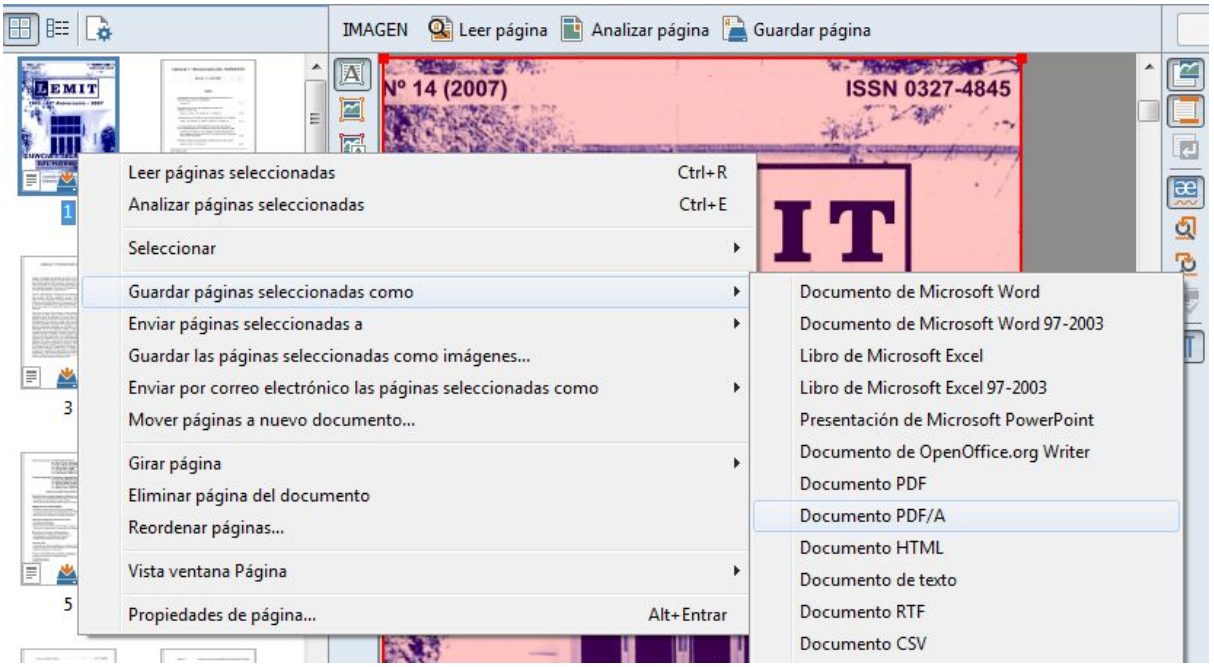
Configuración



Sin compresión, y calidad de imágenes original.

Guardar secciones del documento

Puede realizarse desde Adobe Acrobat professional o desde Abbyy Finereader.



Se utilizara dos procesos de digitalización diferente para los originales y para las copias:

7.2 Tipo de Ficheros y Formatos
Se utilizará formato TIFF para las imágenes máster de cada página con compresión sin pérdida GIII y GIV.
El formato elegido para los archivos derivados es PDF/A, con el OCR correspondiente y optimizados para su descarga en la web.
7.4 Nomenclatura de Ficheros y Carpetas
Se utilizara una nomenclatura descriptiva, utilizando un número consecutivo para la tesis y el apellido del autor sin caracteres especiales, como se muestra en el ejemplo:
Modelo de nombre

Tesis_(no secuencial para cada tesis)_(Apellido del autor)_.(extensión correspondiente)
Másteres:
Tesis_1835_Sanchez (Fichero)
Tesis_1835_Sanchez_page000.tif (cada una de las imágenes TIFF)
Tesis_1835_Sanchez_page001.tif
Tesis_1835_Sanchez_page002.tif

Derivado:
Tesis_1835_Sanchez (Fichero)
Tesis_1835_Sanchez.pdf (archivo en formato pdf para el usuario)