Archivos en el largo plazo

Herramientas para organizar, preservar y digitalizar documentos en sus diversos soportes y formatos

Lunes 13, 20, 27 de septiembre y 4 de octubre

WIKIMEDIA ARGENTINA

La memoria, los trabajos y los días

Equipamiento y procesos de digitalización dentro del repositorio SEDICI

Dra. Marisa R. De Giusti
Lorenzo Calamante
Carlos Nusch
Esteban C. Fernández

UNIVERSIDAD NACIONAL DE LA PLATA

CESGI Centro de Servicios en Gestión de Información

PREBI SEDICI

prebi.unlp.edu.ar sedici.unlp.edu.ar

COMISIÓN DE INVESTIGACIONES CIENTÍFICAS

Ministerio de Ciencia, Tecnología e Innovación

Circuito de digitalización en SEDICI

1. **Recepción, análisis y evaluación del material a digitalizar**
2. **Carga de materiales en el sistema de gestión (Redmine)**
3. **Elección de metodología de escaneo**
4. **Captura de imágenes**
5. **Edición de imágenes**
6. **Guardado de archivos para preservación y difusión**

1) 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:

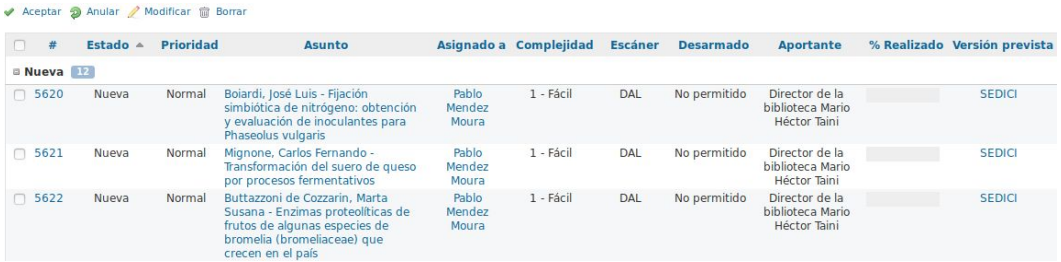
- Estado general de conservación
- Dimensiones
- Formatos
- Tipos de encuadernación
- Importancia histórica, educativa, institucional

2) Carga de materiales en el sistema de gestión (Redmine)

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
- agregan todos los datos propios del material (Autor, Título etc)

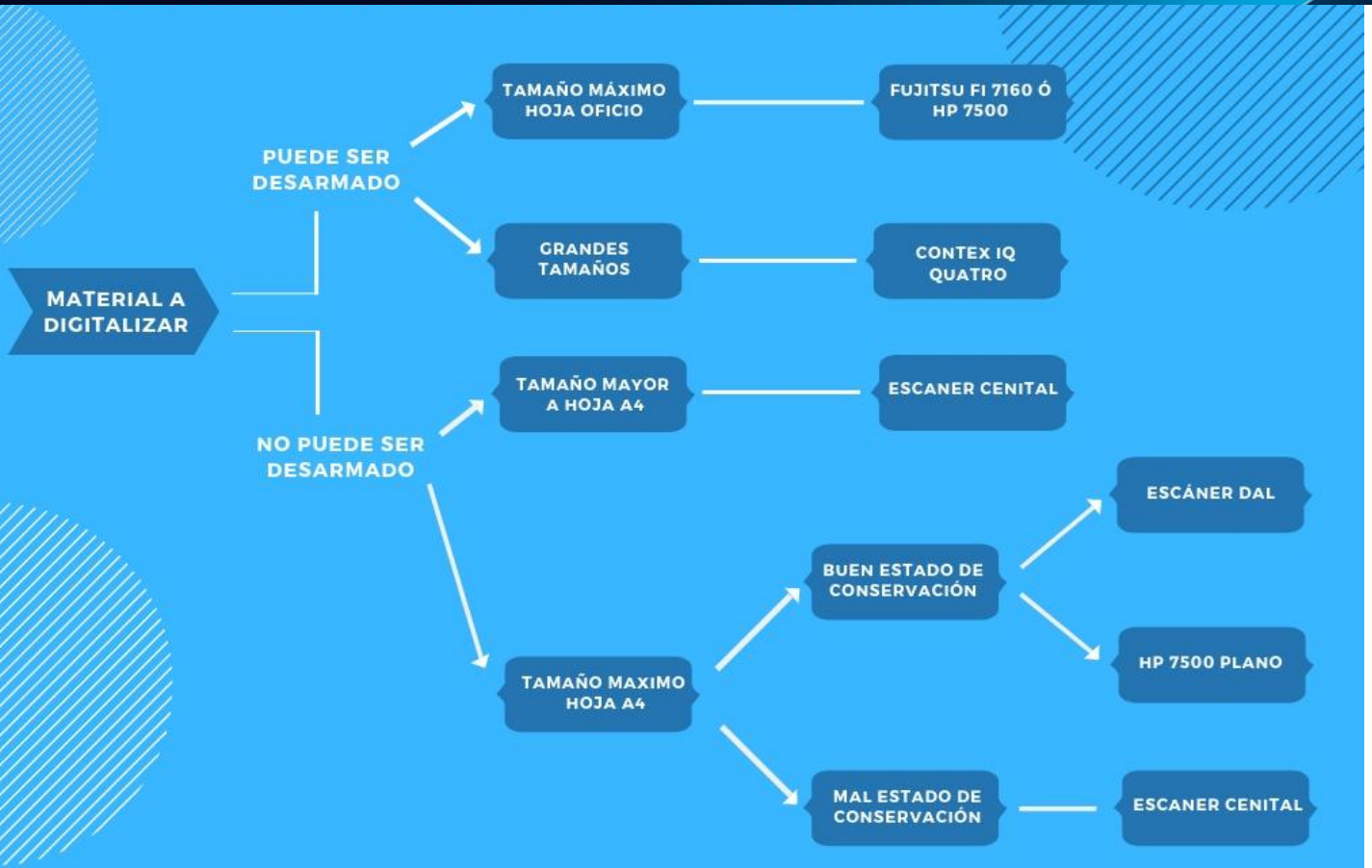
A medida que las obras van pasando por distintas etapas, también se verá reflejado en el sistema hasta que el proceso finaliza.



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

3) Elección de metodología de escaneo

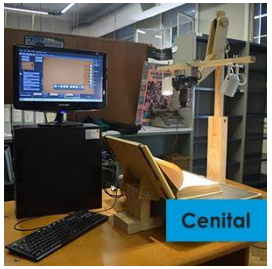


Tipos de escáneres utilizados

automáticos

de gran formato

de libros



Escáneres automáticos

Este tipo de escáner son de alimentación automática, permite así un mayor flujo de material y una mayor velocidad de procesamiento. Además de el alimentador automático el modelo HP 7500 trae una cama plana para digitalizar hojas sueltas, que por distintas razones por ejemplo fragilidad del papel, no pueden ser procesadas a través del alimentador automático.

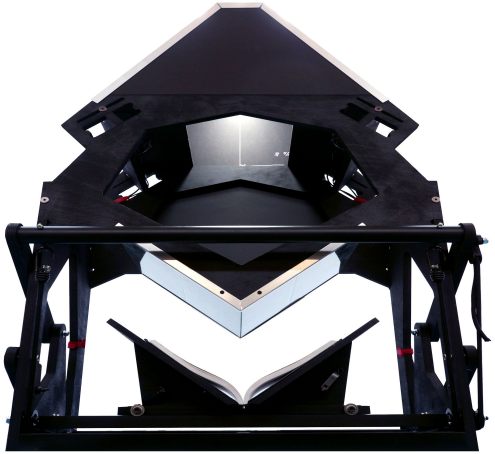
Escáner de gran formato

Permite digitalizar hojas de hasta 44 pulgadas, por ejemplo: mapas, planos, dibujos arquitectónicos, posters, etcétera.

Escáneres de libros

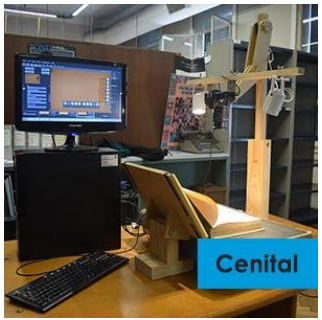
Archivista 2014

Este escáner fue fabricado íntegramente en SEDICI bajo las pautas propuestas por http://diybookscanner.org. Cuenta con dos cámaras Nikon reflex D5300 y es controlado por el software gratuito y de código abierto DigiCamControl http://digicamcontrol.com/



Cenital

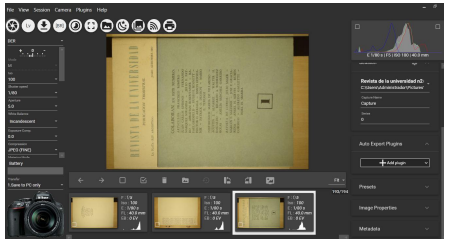
Fue adaptado en SEDICI para digitalizar materiales con dificultades en la manipulación y encuadernaciones frágiles. Por ejemplo las Joyas de la colección Cervantina



4) Captura de imágenes

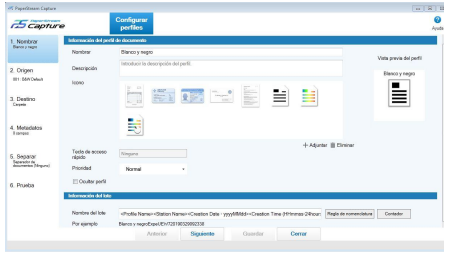
Captura con **“digiCamControl”**

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

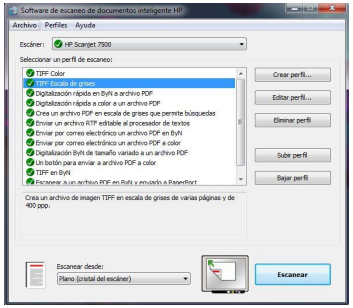


Captura con “**Paperstream**”

Este es el software utilizado para la captura con el escáner Fujitsu FI 7160.

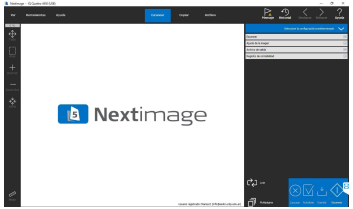


Captura con **“Software de escaneo de documentos inteligente” de HP**



Este programa es utilizado para controlar el escáner HP 7500

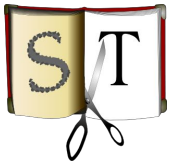
Captura con **“NextImage”**



Este software es utilizado para controlar el escáner de formato grande Contex IQ Quattro.

5) Edición de imagen

Luego de obtener las imágenes escaneadas, se editan en Photoshop o ScanTailor. De esta manera se puede corregir la orientación, dividir y alinear las páginas, seleccionar el contenido, limpiar los márgenes, eliminar las manchas y modificar el contraste.



Edición con Photoshop

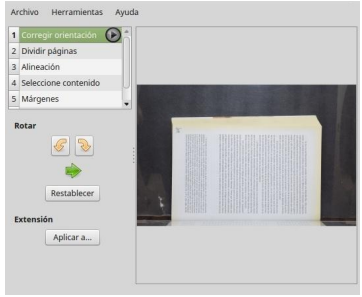
Este programa es uno de los más potentes del mercado para la edición de imágenes, y es utilizado en casos que presentan muchas dificultades en la visualización o legibilidad.

Edición con ScanTailor Advanced

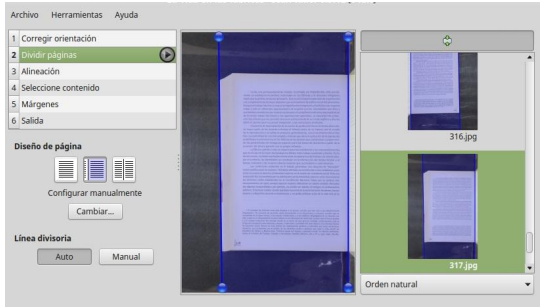
Scantailor es una herramienta gratuita de código abierto que permite corregir o modificar las imágenes capturadas. Soporta los siguientes formatos de entrada: *.tif, *.tiff, *.png, *.jpg, *.jpeg y genera archivos con formato tiff de salida (uno por cada página).

Algunas de las funciones principales de ScanTailor son las de:

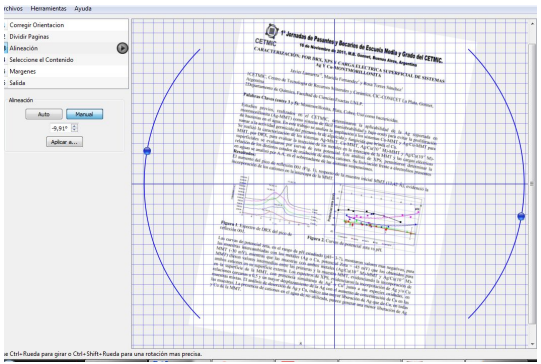
Rotar páginas

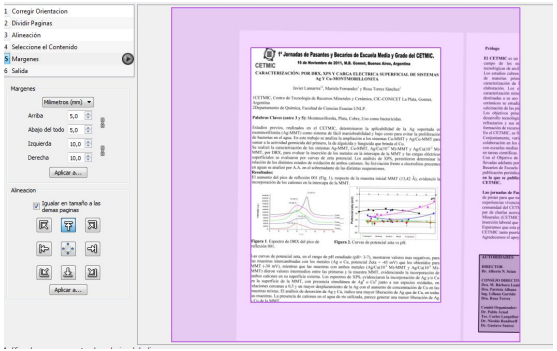


Dividir páginas

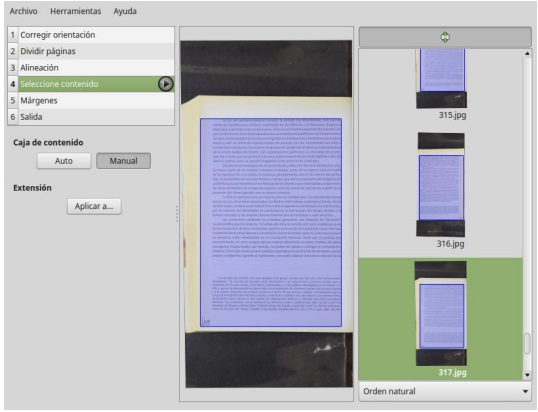


También es posible corregir la orientación de las páginas, seleccionar los márgenes, eliminar manchas y por último ajustar el color y contraste de las páginas.

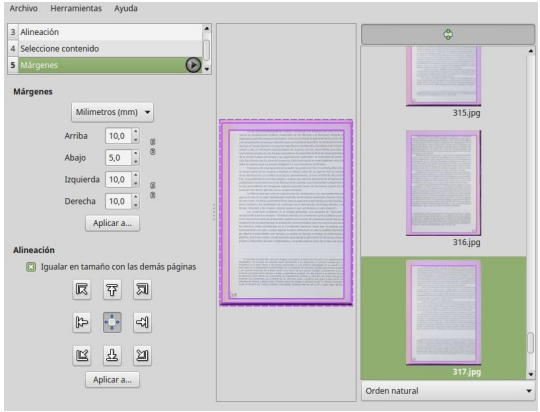




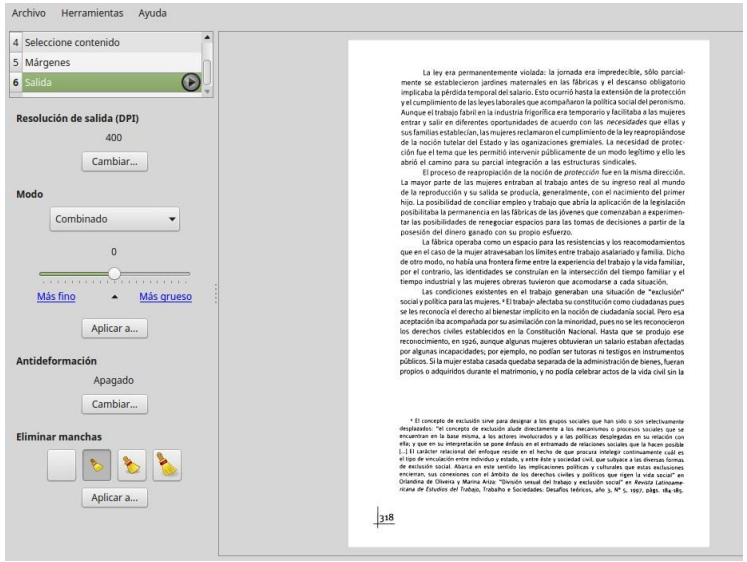
Seleccionar contenido



Agregar márgenes



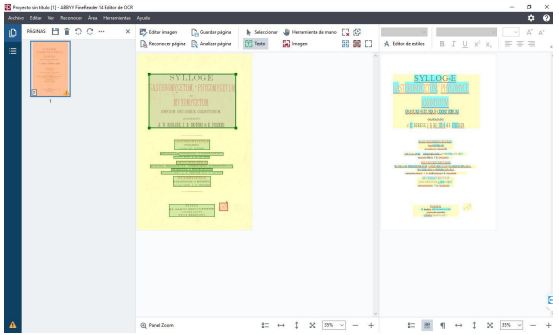
Eliminar manchas



6) Guardado de archivos para preservación y difusión

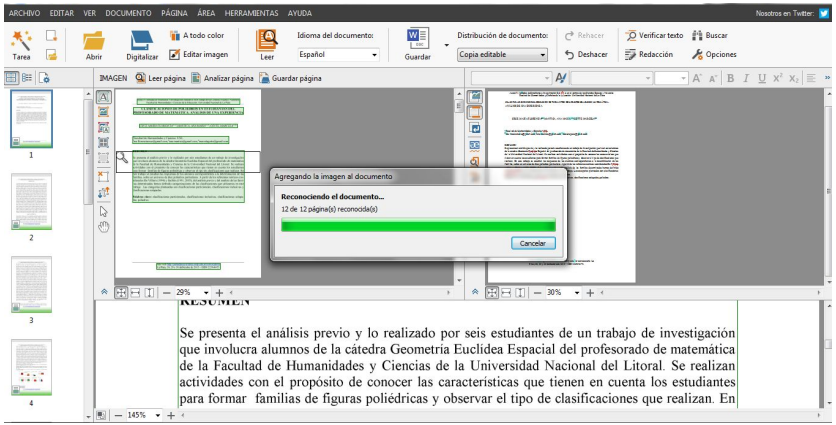
Reconocimiento de caracteres (OCR) y guardado en formato PDFA con Abbyy FineReader

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

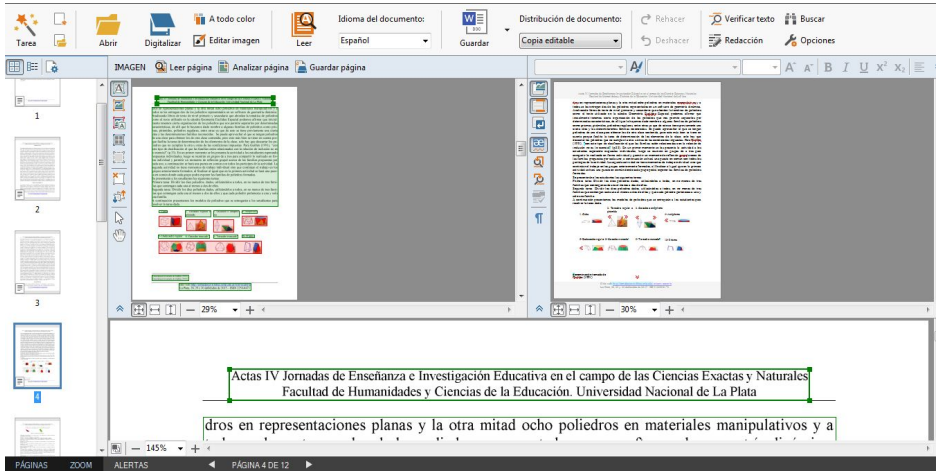


Generación de OCR con ABBYY FineReader

ABBYY FineReader es un software de OCR que permite trabajar y editar pdf de manera rápida y confiable.



Este software permite seleccionar imágenes e indicar qué parte de la página debe ser reconocido y qué no. Posee un motor de reconocimiento óptico de caracteres muy potente además también permite la corrección manual.



Preservación digital

La preservación digital se define como el conjunto de prácticas de naturaleza política, estratégica y acciones concretas, destinadas a asegurar la preservación, el acceso y la legibilidad de los objetos digitales a largo plazo.

Una estrategia de preservación es la de adoptar estándares internacionales, es decir, apoyarse en la afirmación de que los estándares internacionales son relativamente estables en el tiempo.

Según las guias "Technical Guidelines for Digitizing Cultural Heritage Materials" (FADGI), los formatos de archivos utilizados para preservación son: **TIFF, JPEG2000 y PDF/A.**

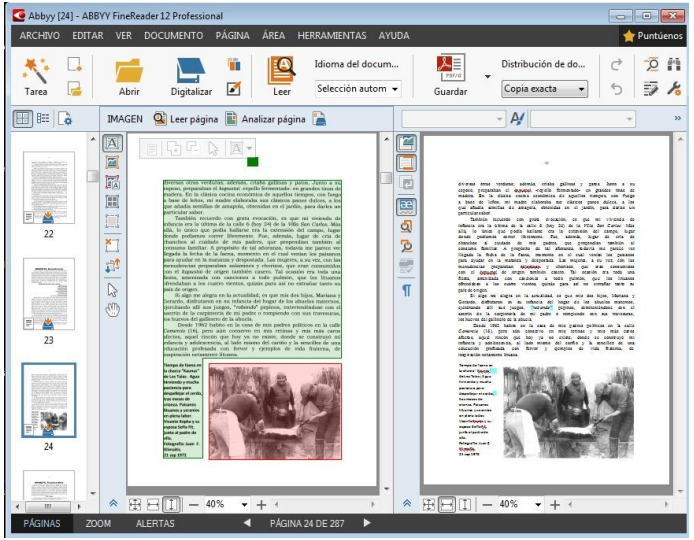
Dentro del repositorio SEDICI utilizamos el formato TIFF para el guardado de archivos maestros y PDF/A para archivos de preservación y difusión.

PDF/A (Portable Document Format) es uno de los mejores formatos para preservar documentos electrónicos. Se utiliza la versión /A (Archival) para archivar documentos con fines de preservación, pues contiene todos los elementos necesarios para reproducir el contenido tal como se generó, independientemente del programa con que se creó.

TIFF (Tagged Image File Format) es un formato de imágenes muy usado y de estándar abierto. Los archivos pueden utilizar compresión sin pérdidas y es utilizado para la creación de archivos maestros de imagen.

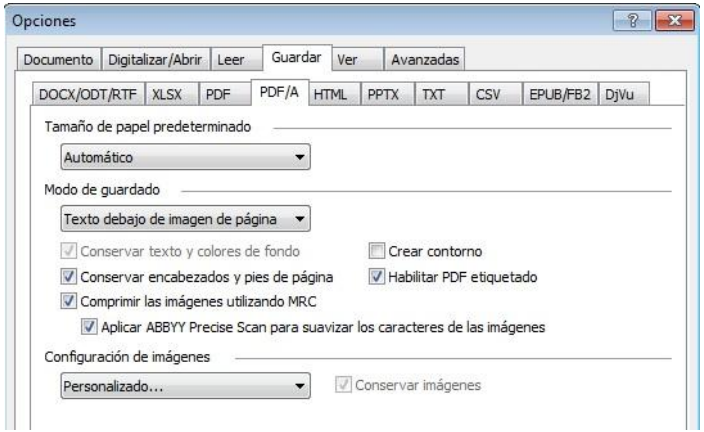
Reconocimiento de caracteres (OCR)

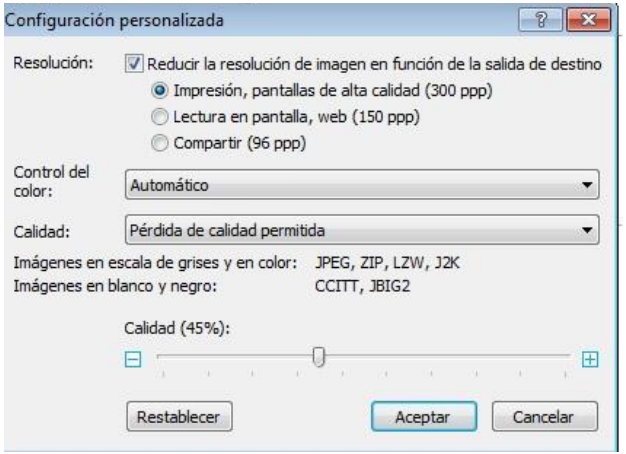
Luego de editar las imágenes se realiza el OCR con el Abbyy FineReader. En esta etapa del proceso se selecciona el contenido según sea texto, imagen o cuadro. Luego se revisa el resultado del OCR y se generan los archivos PDF/A.



Compresión de pdf

Por último, en el momento del guardado, el programa nos permite modificar la compresión para obtener documentos más pequeños, que pueden ir desde compresiones sin pérdida a compresiones con pérdida de calidad.

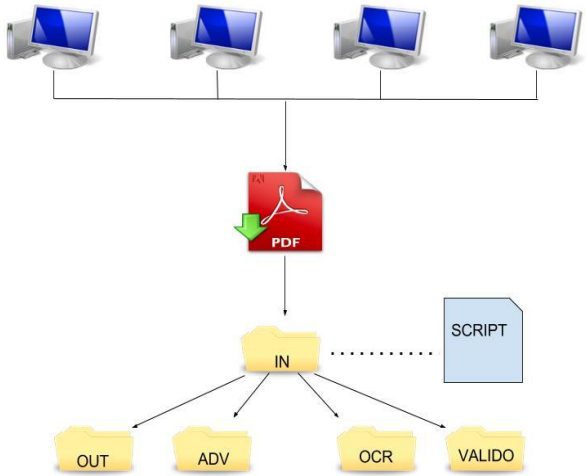




3-HEIGHT

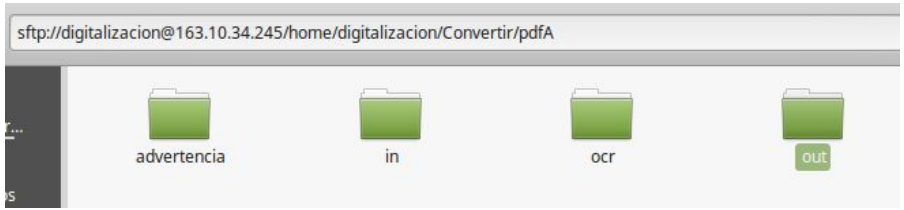
Este software posee una arquitectura cliente servidor, que permite convertir por lotes archivos de distintos formatos a pdf/a. Además también es utilizado para verificar si los archivos pdf/a cumplen con la norma.

- Deteccion de archivos
- Análisis
- Conversión
- Verificación



Simplemente tenemos una carpeta compartida con el nombre PDFA que consta de 4 directorios donde los administradores podrán transformar los archivos PDF en PDFA. Los directorios son:

- Una carpeta “in” para ingresar los archivos a procesar
- Una Carpeta “out” donde se depositarán los archivos resultantes.
- Y dos carpetas destinadas a diferentes tipos de errores llamadas “advertencia” y “ocr”



Generación de PDF/A con 3-HEIGHT

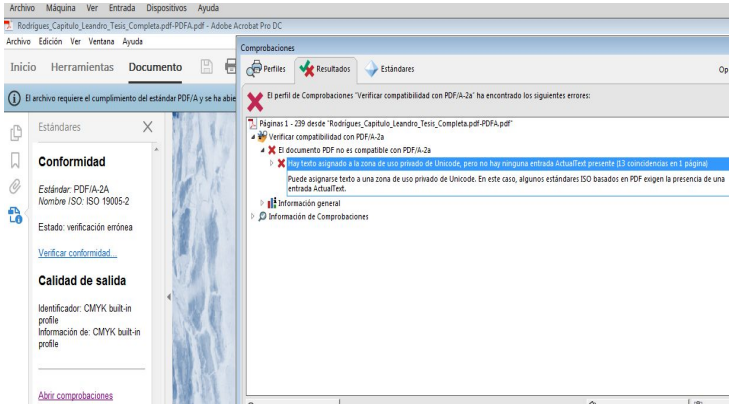
El 3-HEIGHT analiza el pdf y elige en qué versión va a convertirlo. Si la conversión sale bien, en la carpeta out tendremos los siguientes archivos:

El archivo con la fecha 13-20-16.txt presenta el log de la ejecución del script..
El archivo pdf original.
El archivo convertido con la terminación: -PDFA.pdf
El último archivo txt da más detalles de la conversión del archivo original

Validación de PDF/A - Acrobat DC

Una vez obtenido el pdf/A de 3-Height es necesario validarlo también en Acrobat DC. Si la verificación es errónea dependiendo el caso de error podemos arreglarlo desde el mismo Acrobat. Por ejemplo: cuando un archivo no pasa la verificación porque el texto no es unicode en todo el pdf. Generalmente este problema se soluciona transformando el archivo en la versión de pdfA llamada pdf/A-2u.

Adobe Acrobat DC



Generación de PDF/A

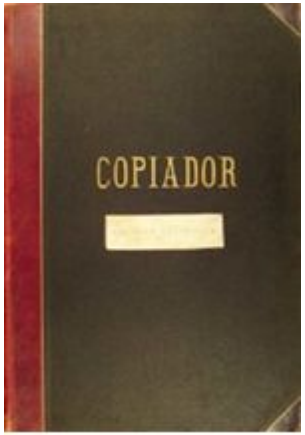
Un formato de preservación para documentos de texto es el estándar PDF/A, descrito en las normas ISO 19005-(1-2-3). Está basado en el estándar PDF 1.4, al que le incorpora algunos requerimientos adicionales, por ejemplo:

- Especificaciones sobre los metadatos y la estructura del archivo.
- La paleta de colores (incluyendo escala de grises y blanco/negro) no deben ser representados en un espacio de color de dispositivo (DeviceRGB, DeviceCMYK, DeviceGray).
- Las fuentes usadas en texto visibles deben estar embebidas (incluídas dentro del archivo).

Uno de los propósitos de los requerimientos del estándar PDF/A es de proveer soporte para personas con capacidades diferentes, por ejemplo, incorporando la información requerida y necesaria para aplicaciones que hagan el pasaje de texto a voz.

Ejemplo de caso de proceso completo de digitalización:

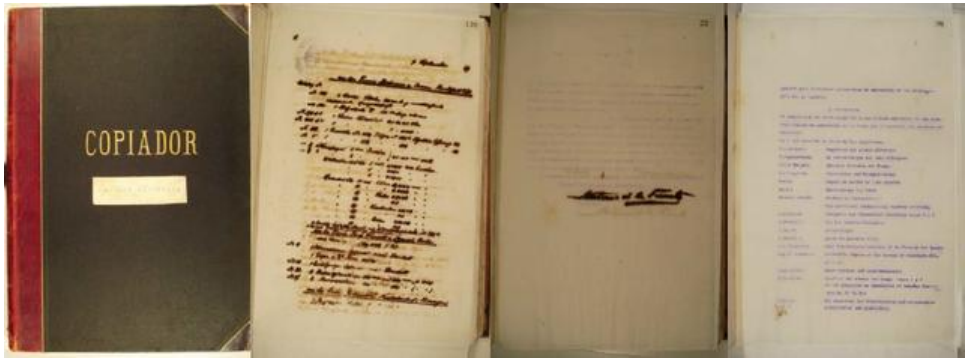
LIBRO COPIADOR - FACULTAD DE CS. FÍSICAS, MATEMÁTICAS Y ASTRONÓMICAS (1918-1925)



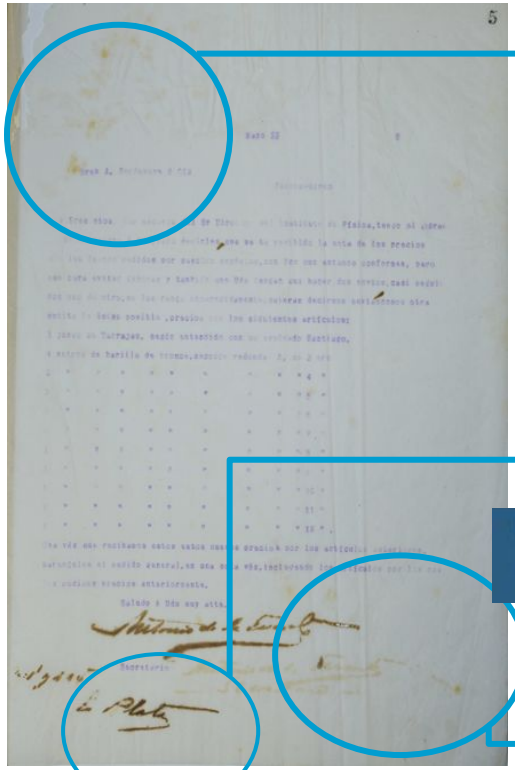
SEDICI
SERVICIO DE DIFUSIÓN DE LA CREACIÓN INTELECTUAL

SEDICI y el **Museo de Física de la Facultad de Ciencias Exactas** de la **UNLP** destinaron personal para la digitalización de un documento archivístico: el libro *Copiador – Facultad de Ciencias Físicas, Matemáticas y Astronómicas (1918-1925).* Se siguieron los estándares internacionales para la digitalización (IFLA, NARA, FADGI, etc.), pero **muchas de las dificultades que presentó el material no estaban contempladas en la bibliografía**.

Estado de conservación

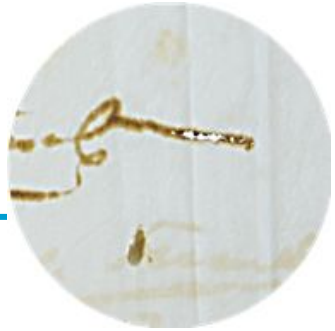


- Papel amarronado, débil, friable con roturas y desprendimientos.
- Escritura manuscrita con tinta difundida en el papel y transferida a los siguientes, con pérdida de nitidez e imagen doble.
- Escritura mecanográfica poco legible.
- Encuadernación con especiales requerimientos de manipulación.



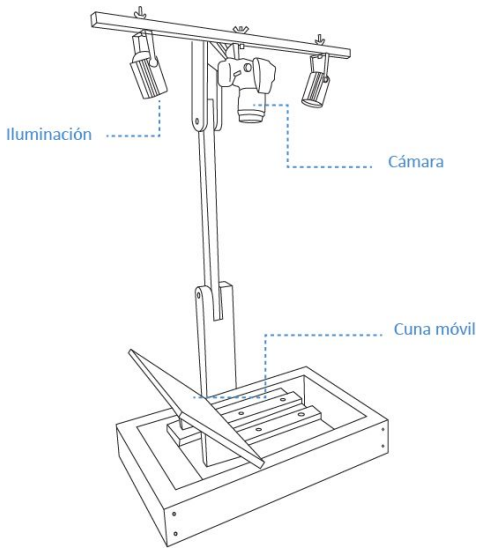
Dobleces y desprendimientos

Escritura mecanografiada poco legible y pérdida de nitidez



Tinta difundida en el papel y transferida a los consecutivos.

Escáner adaptado para este propósito:



Se optó por un sistema de escaneo **rediseñado a partir del Model 1** de DIY, con una cámara cenital apuntando hacia el libro, junto con dos luces LED dicroicas de luz cálida cuya temperatura no daña el material.

Software de captura

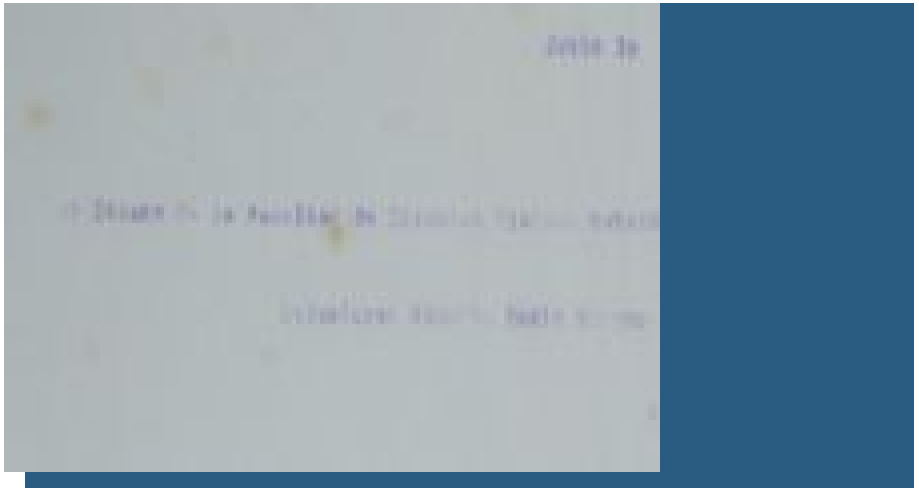
El **digiCamControl 2.0.72.0** fue un programa rápido, confiable y versátil para la captura de imágenes y permitió el manejo de las cámaras directamente desde la computadora.

Software de edición

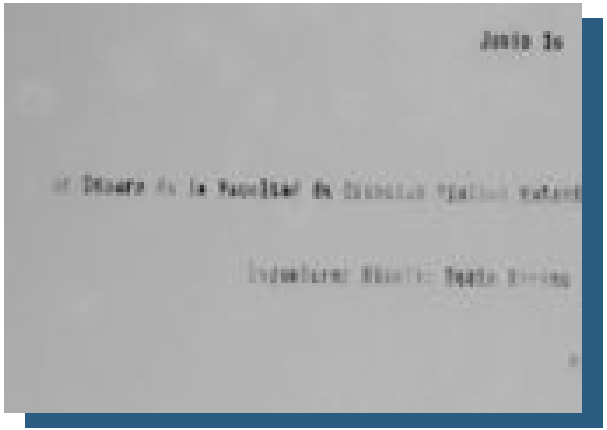
Photoshop CS6 permitió mayor libertad en la manipulación de las imágenes destinadas al reconocimiento de texto. Se aplicaron filtros y se automatizó el procedimiento estándar para todas las imágenes

Post-procesos de ajuste de imagen y enfoque (Photoshop)

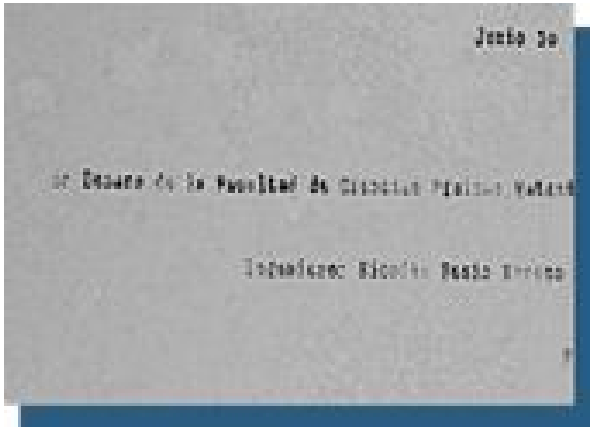
Se utilizaron dos filtros para mejorar la imagen capturada con el fin de hacer el OCR del documento



- **Desaturación por color (black and white filter):** este filtro desatura los colores por separado. Esto permite seleccionar las tonalidades que representan manchas, suciedades y atenuarlos hasta que la superficie se vea homogénea.

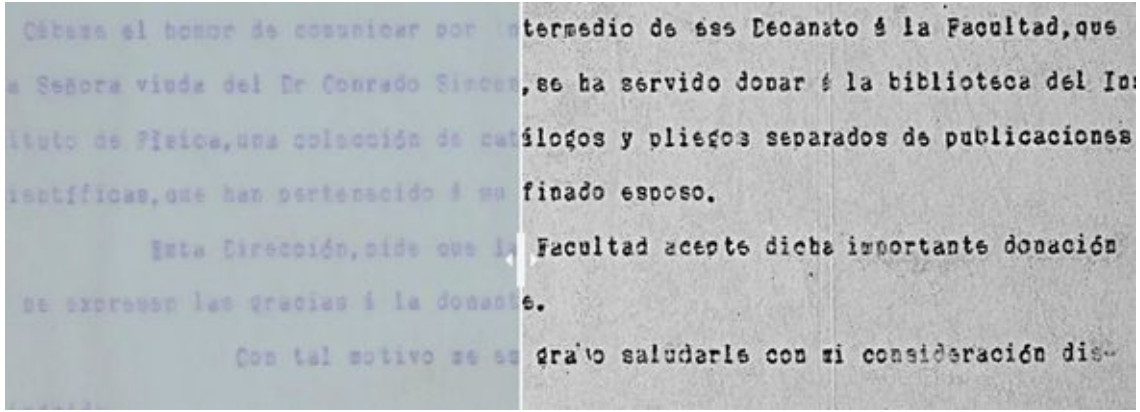


- **Enfocar (smart sharpen)** para acentuar el borde de la tipografía en la imágen y mejorar el contraste con el fondo.



El proceso completo se automatizó completamente por medio de las funciones ***Actions*** y ***Droplet*** de Photoshop

Imagen original e imagen mejorada lista para reconocimiento de texto (Abbyy FineReader)



Por consultas: marisa.degiusti@sedici.unlp.edu.ar

Nuestros sitios

- http://sedici.unlp.edu.ar
- http://digital.cic.gba.gob.ar/
- http://cesgi.cic.gba.gob.ar/
- http://prebi.unlp.edu.ar
- http://www.istec.org/liblink/
- http://revistas.unlp.edu.ar/cientificas/
- http://revistas.unlp.edu.ar
- http://congresos.unlp.edu.ar
- http://ibros.unlp.edu.ar