ARCHIVO GENERAL DE LA NACIÓN COLOMBIA

MINCULTURA

TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

Manual de Usuario

SOFTWARE

gimageReader 3.2.99

Grupo de Innovación y apropiación de Tecnologías de la Información Archivística

Compilador: Pedro Antonio Gómez Guarín

2018

Introducción

gImageReader es una herramienta diseñada para el reconocimiento y la extracción de los caracteres contenidos en un documento ya sea en formato PDF o una imagen, mediante la técnica OCR (siglas en inglés para Reconocimiento Óptico de Caracteres).

gImageReader es una interfaz gráfica para la herramienta de software libre Tesseract-OCR, que es un motor de Reconocimiento Óptico de Caracteres (OCR) Inicialmente desarrollado por HP Labs como software propietario, fue publicado como código abierto en el año 2005 y desde el año 2006, Google se ha hecho cargo de su desarrollo y está disponible para múltiples sistemas operativos.

Como principales características este programa puede:

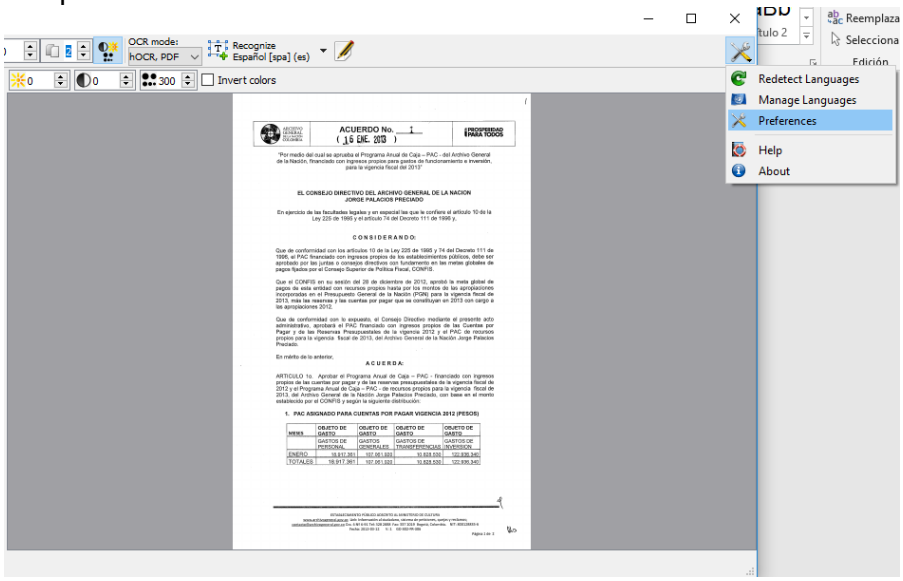
- Importar documentos e imágenes PDF desde el disco, dispositivos de escaneo, portapapeles y capturas de pantalla
- Procesar imágenes y documentos múltiples de una sola acción
- Definir el área de reconocimiento manual o automático en el documento
- Reconocer a texto plano o a documentos hOCR
- Mostrar el texto reconocido directamente al lado de la imagen que se está procesando
- Postprocesamiento del texto reconocido incluyendo la revisión ortográfica
- Genera documentos PDF a partir de documentos hOCR

gImageReader fue utilizado como herramienta para la extracción del contenido de los documentos escaneados en PDF del repositorio normativo del Archivo General de la Nación.
http://normativa.archivogeneral.gov.co/

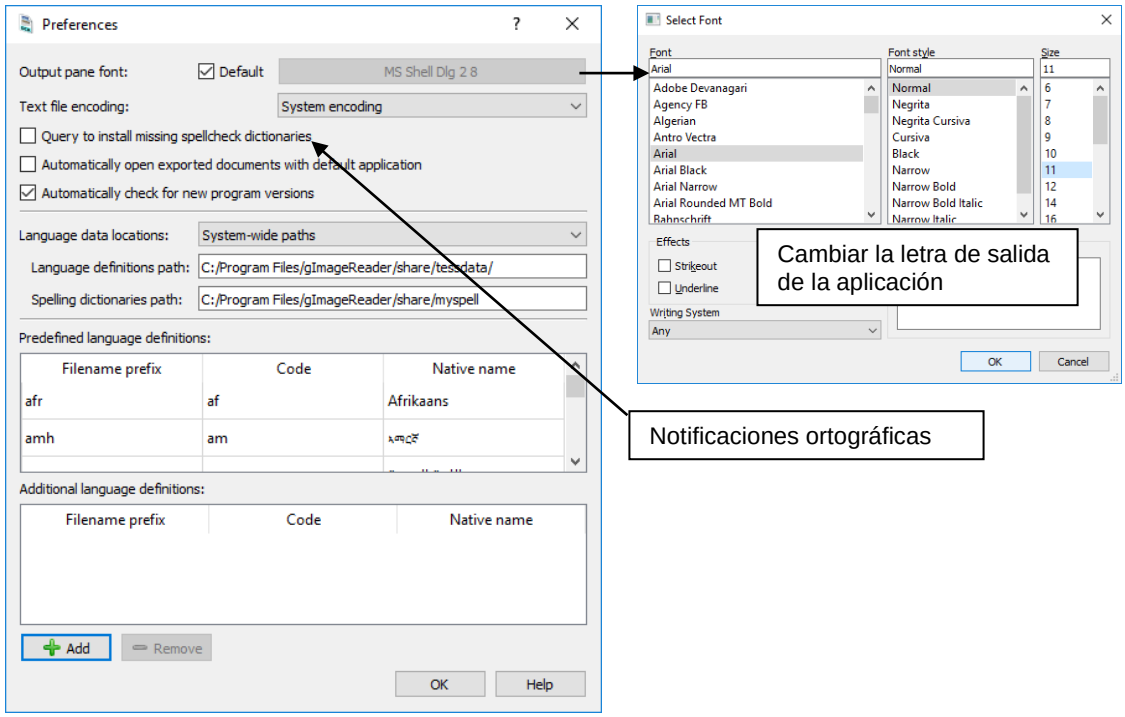
Para realizar este manual utilizamos la versión de gImageReader 3.2.99 para Windows

Opciones de programa

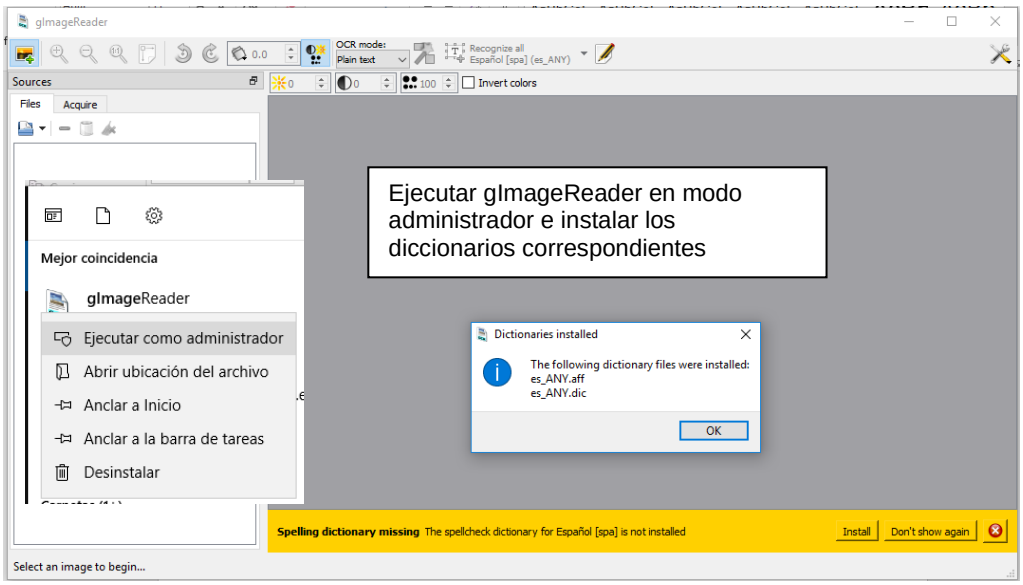
- Se puede acceder a las opciones del programa desde el menú de la aplicación, que se abre al hacer clic en el botón situado más a la derecha de la barra de herramientas principal. Al ejecutar la aplicación.



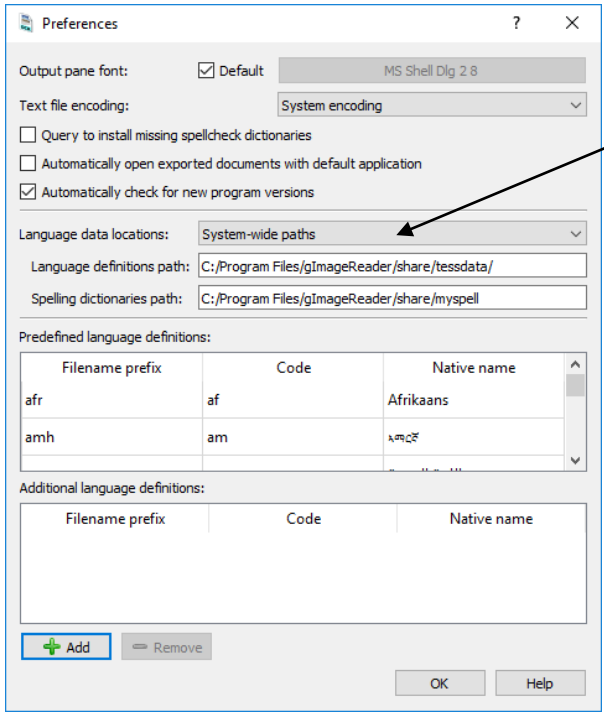
- Las opciones permiten configurar la fuente del panel de salida, así como determinar si la aplicación notificará sobre los diccionarios ortográficos faltantes y las nuevas versiones del programa.



- Al activar las notificaciones ortográficas, es posible que se deba instalar el diccionario para español, se recomienda ejecutar gImageReader en modo administrador.



- La configuración de ubicación de datos del idioma permite controlar si las definiciones del lenguaje tesseract y los diccionarios de ortografía se guardan en todo el sistema (es decir, % ProgramFiles% en Windows o por debajo de / usr en Linux) o en el directorio local del usuario (es decir, esto es útil si el usuario no tiene privilegios de escritura en ubicaciones de todo el sistema.

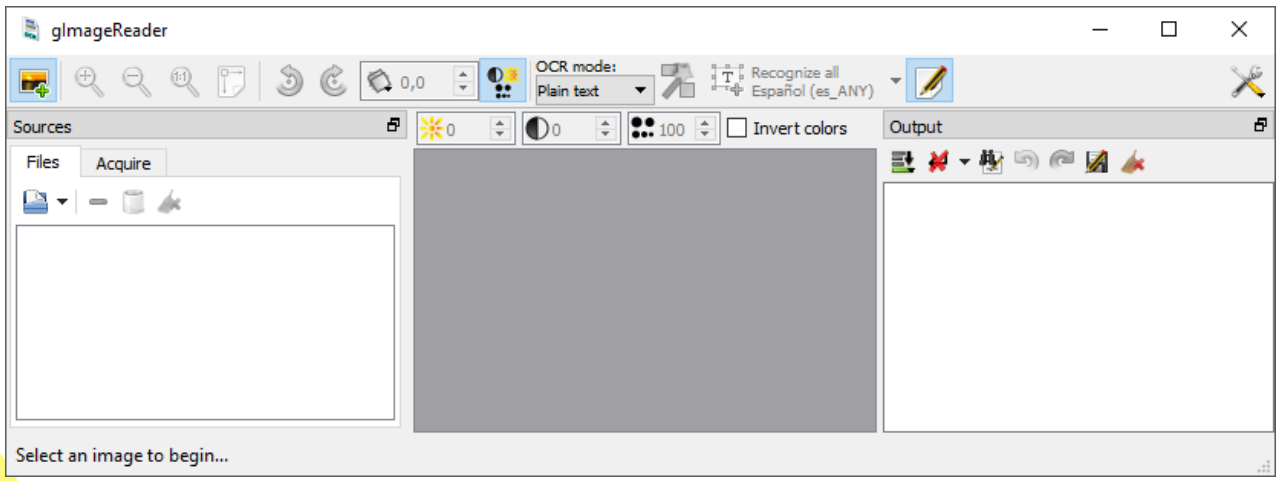


Se puede definir si la instalación de los diccionarios ortográficos se hara en el sistema de archivos del Sistema operativo o en las capetas de usuario local

Principales ítems de la interfaz gImageReader

El proceso de reconocimiento de texto de gImageReader tiene ciertos pasos o instancias las cuales permiten realizar un perfecto reconocimiento del texto de una imagen.

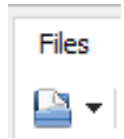
A continuación, explicaremos estos pasos detalladamente.



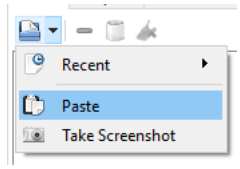
Apertura e importación de imágenes

- Las imágenes se pueden abrir / importar desde el panel de fuentes, que se activa al hacer clic en el botón superior izquierdo de la barra de herramientas principal.

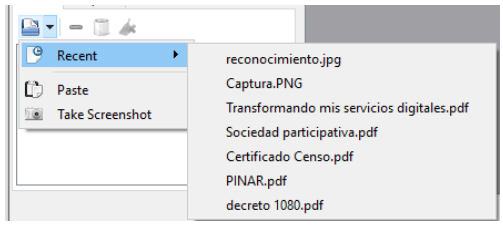
- Para abrir una imagen existente o documento PDF, haga clic en el botón Abrir en la pestaña de imágenes.



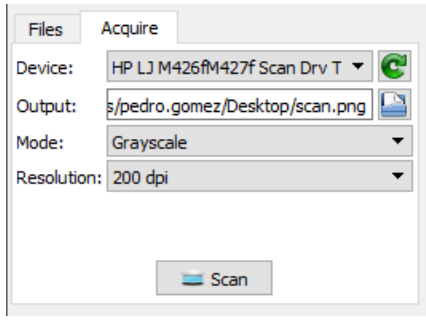
- Para realizar una captura de pantalla, pegar datos de imágenes del portapapeles o abrir un archivo recientemente abierto, haga clic en la flecha al lado del botón Abrir.



- Puede administrar la lista de imágenes abiertas con los botones al lado del botón Abrir.

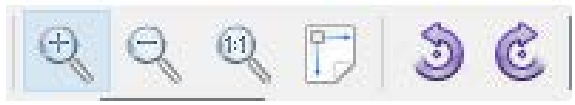


- Para adquirir una imagen desde un dispositivo de escaneo, haga clic en la pestaña de adquisición en el panel de fuentes.



Ver y ajustar imágenes

- Use los botones en la barra de herramientas principal para acercar y alejar, así como para rotar la imagen en un ángulo arbitrario.

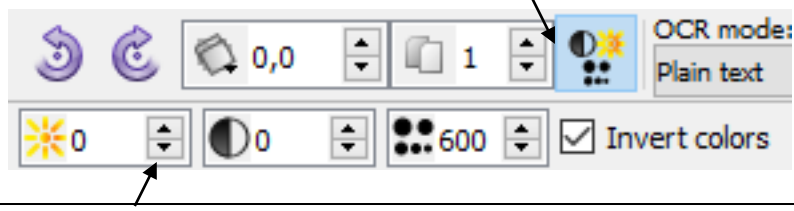


- El zoom también se puede realizar desplazándose sobre la imagen con la tecla CTRL presionada.

- Las herramientas básicas de manipulación de imágenes se proporcionan en la barra de herramientas de controles de imagen, que se activa haciendo clic en el botón de controles de imagen en la barra de herramientas principal. Las herramientas proporcionadas actualmente permiten ajustes de brillo y contraste, así como también ajustan la resolución de la imagen (a través de la interpolación).

Para un mejor resultado se recomienda utilizar una resolución de 600 y el inversor de color

botón de controles de imagen



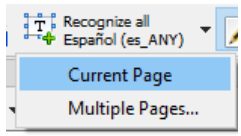
Los controles de imagen permiten ajustar parámetros teles como la luminosidad, el contraste la resolución y de ser necesario invertir los colores para ayudar al motor de reconocimiento a hacer mejor el trabajo

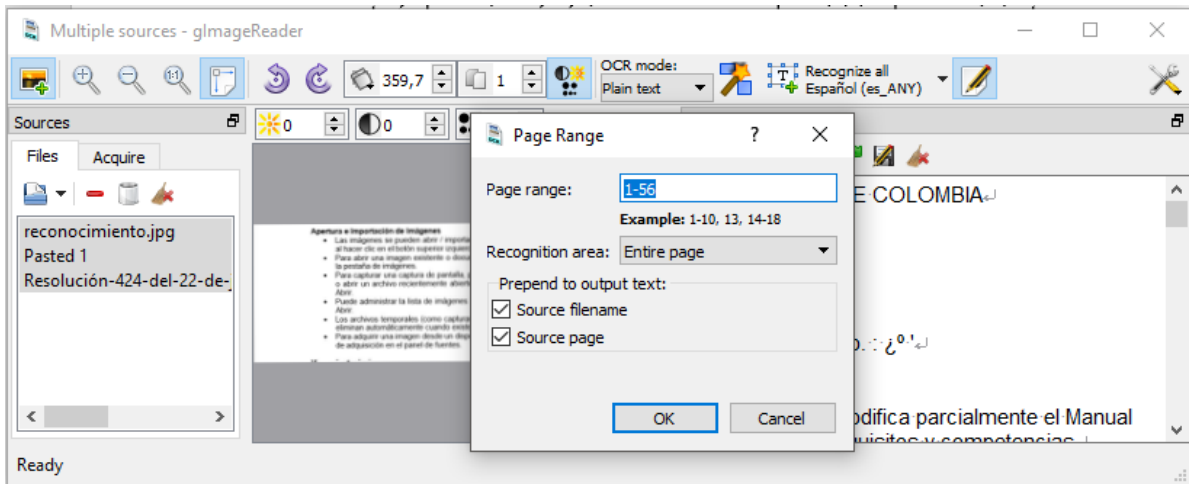
Preparándose para el reconocimiento

- Para realizar OCR en una imagen, el usuario necesita especificar:
  - Las imágenes de entrada (por ejemplo, imágenes para reconocer),
  - El modo de reconocimiento (por ejemplo, texto plano vs hOCR, PDF)
  - El idioma de reconocimiento.

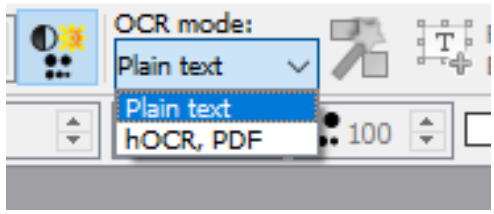
- Las **imágenes de entrada** corresponden a las entradas seleccionadas en la pestaña de imágenes en el panel de fuentes. Si se seleccionan varias imágenes, el programa tratará el conjunto de imágenes como documento de varias páginas y preguntará al usuario qué páginas procesar cuando se inicie el reconocimiento.

El botón reconocer al español despliega dos opciones, Pagina actual (Current page) o Múltiples paginas (Multiple pages)

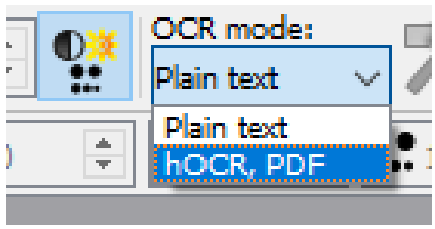


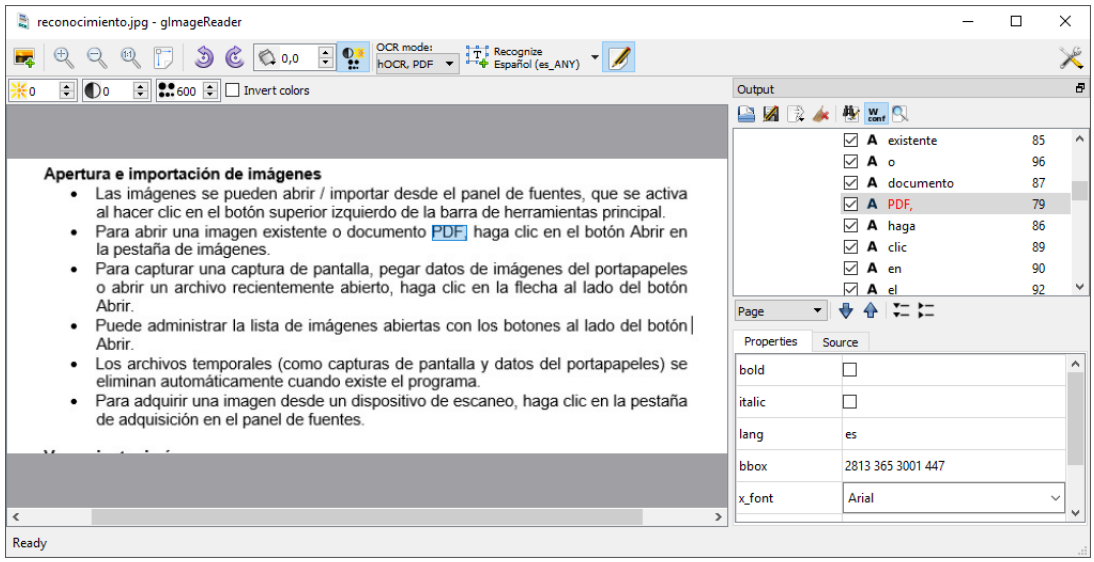


- El **modo de reconocimiento** se puede seleccionar en el cuadro combinado de modo OCR en la barra de herramientas principal:
    - El modo de texto sin formato hace que el motor de OCR extraiga solo el texto sin formato, sin información de formato y diseño.

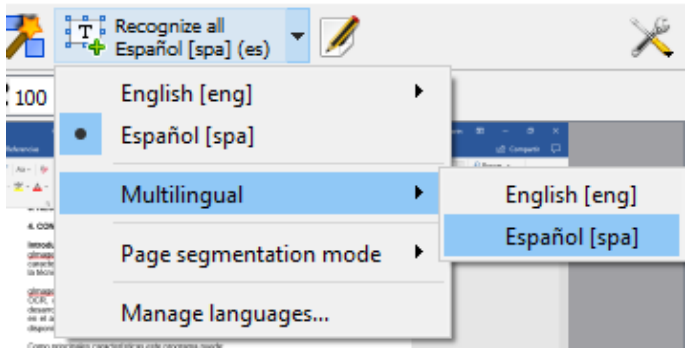


    - El modo hOCR, PDF hace que el motor de OCR devuelva el texto reconocido como un documento htm html, que incluye información de formato y diseño para la página reconocida. hOCR es un formato estándar para almacenar los resultados de reconocimiento y se puede utilizar para interactuar con otras aplicaciones compatibles con este estándar. gImageReader puede procesar HOCR ^ documentos además para generar un documento PDF para el resultado del reconocimiento.

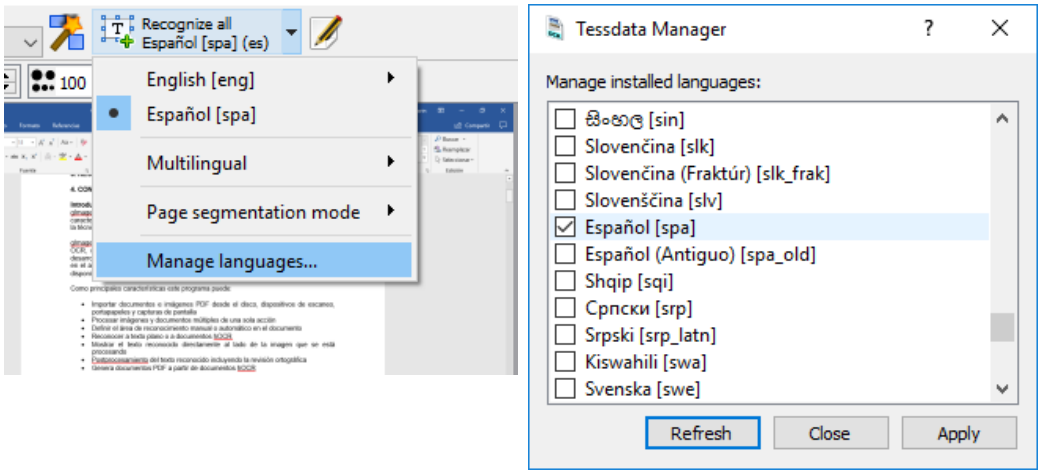




- El **idioma de reconocimiento** se puede seleccionar desde el menú desplegable del botón de reconocimiento en la barra de herramientas principal, instalando un diccionario de ortografía para una definición de lenguaje tesseract, es posible elegir entre las diferentes regiones disponibles de idioma, esto solo afectará el idioma para la revisión de la ortografía del texto reconocido.

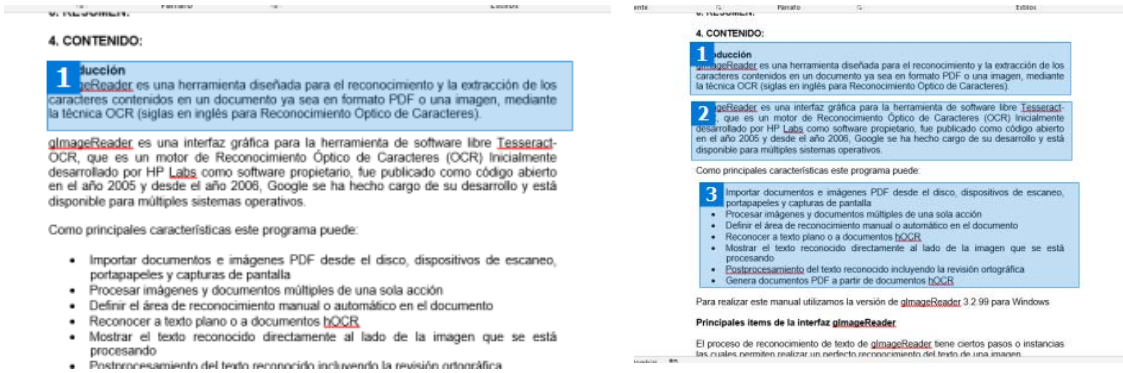


Se pueden especificar múltiples idiomas de reconocimiento a la vez desde el submenú Multilingüe del menú desplegable. Las definiciones del lenguaje tesseract instalado se pueden administrar desde Administrar idiomas e instalar el idioma necesario.

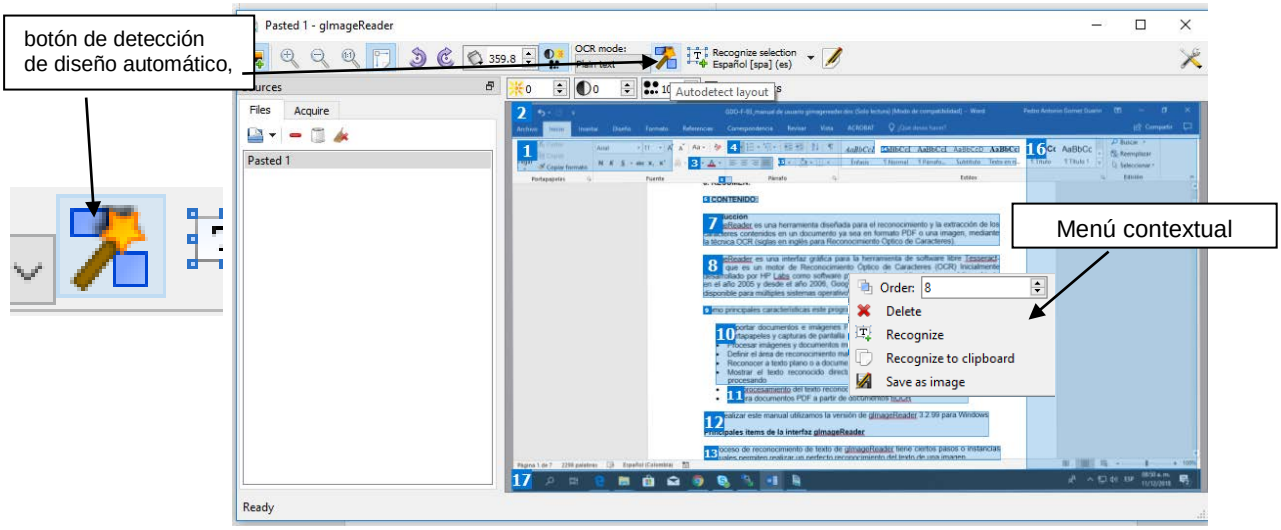


Reconocimiento y post-procesamiento en modo de texto sin formato

- Las áreas que deben reconocerse se pueden seleccionar arrastrando **(clic con el botón izquierdo + movimiento del mouse)** un área rectangular alrededor de las partes de la imagen. Es posible realizar varias selecciones presionando la tecla CTRL mientras selecciona el texto.



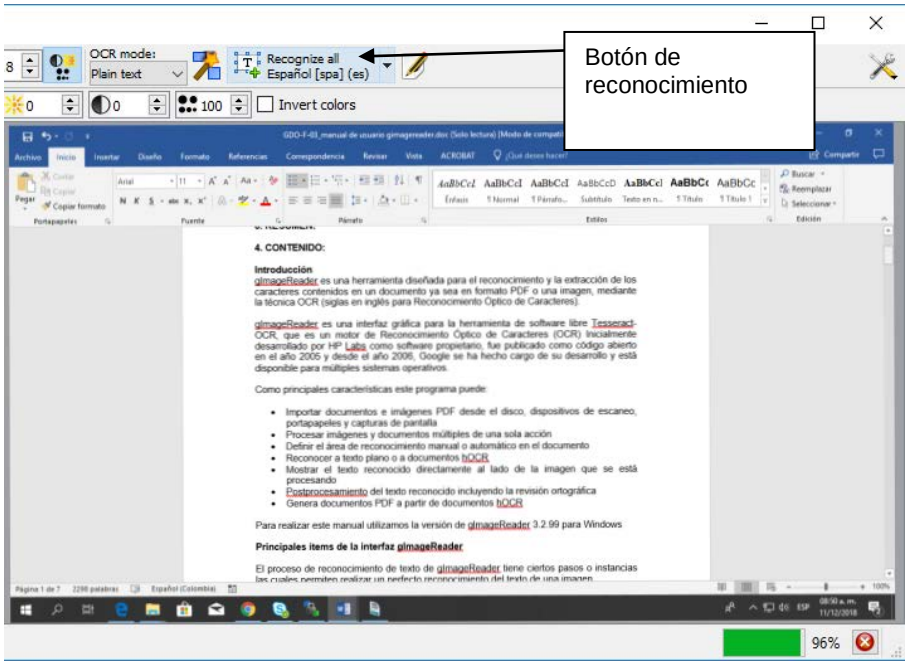
- Alternativamente, el botón de detección de diseño automático, accesible desde la barra de herramientas principal, intentará definir automáticamente las áreas de reconocimiento apropiadas, así como ajustar la rotación de la imagen si es necesario.



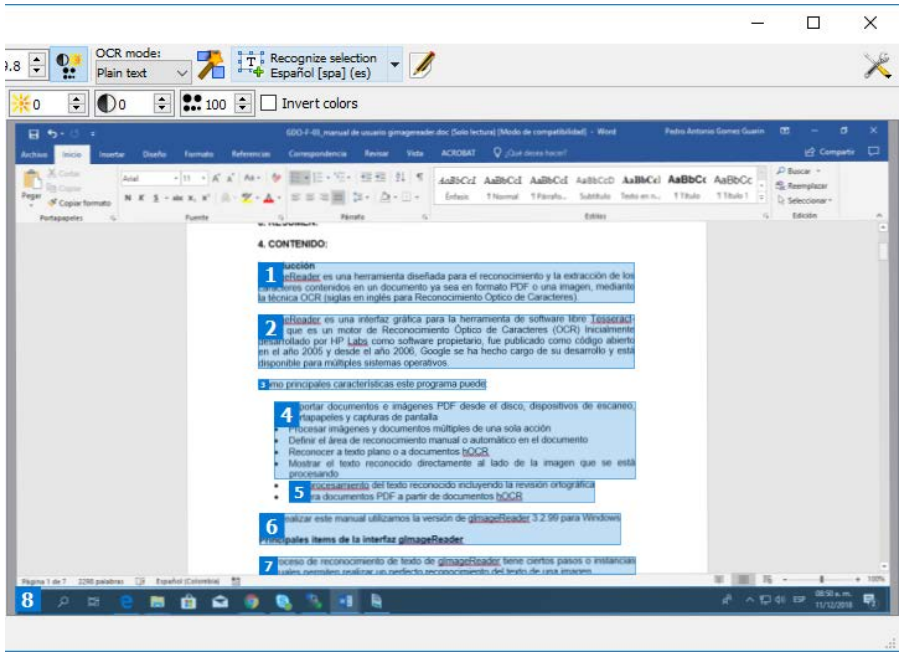
- Las selecciones se pueden eliminar y reordenar a través del menú contextual que aparece al hacer clic derecho sobre ellas. También es posible cambiar el tamaño de las selecciones existentes arrastrando las esquinas del rectángulo de selección.

- Las porciones seleccionadas de la imagen (o la imagen completa, si no se definen las selecciones) se pueden reconocer al presionar el botón de reconocimiento en la barra de herramientas principal. Alternativamente, las áreas individuales pueden reconocerse haciendo clic derecho en una selección. Desde el menú contextual de selección, también es posible redirigir el texto reconocido al portapapeles, en lugar del panel de salida.

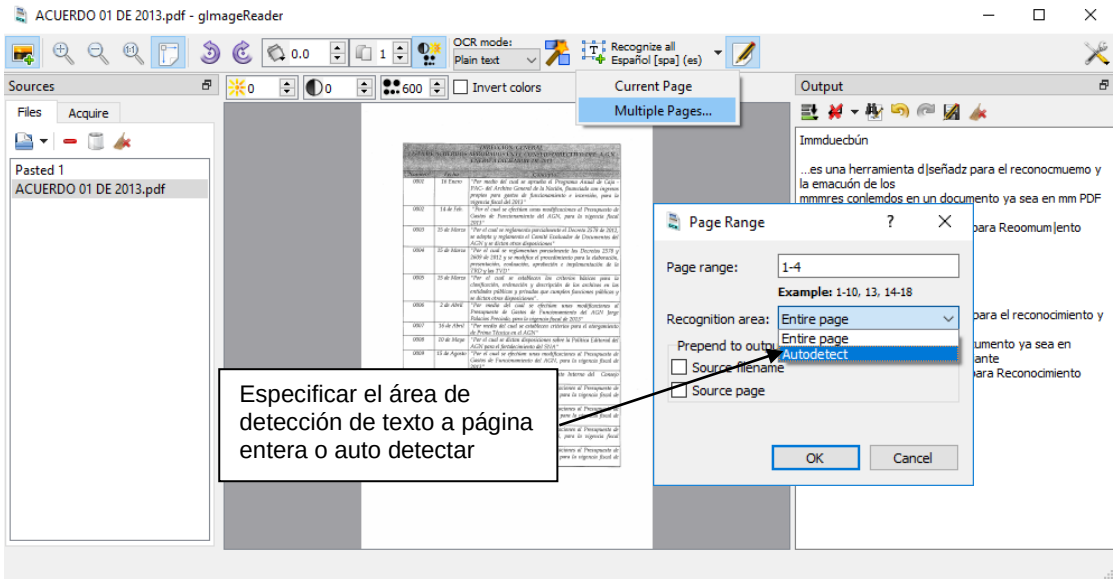
Reconocimiento sin selección



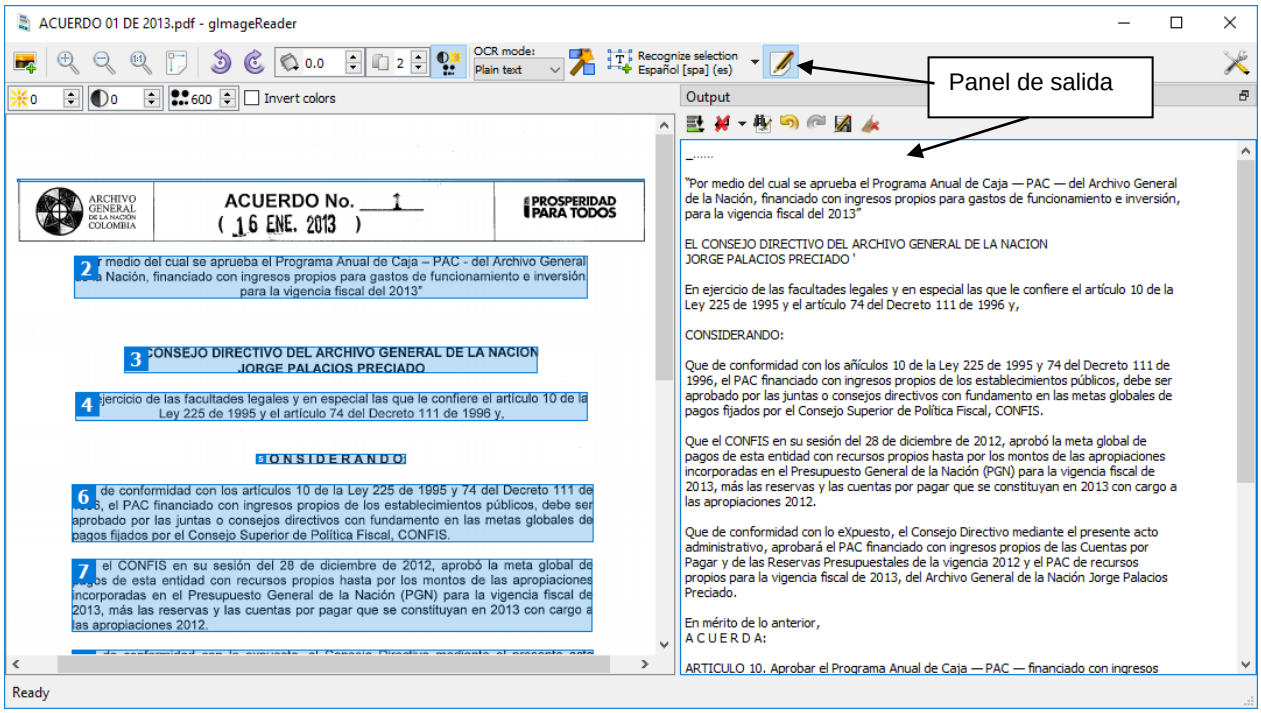
Reconocimiento con selección



- Si se seleccionan varias páginas para reconocimiento, el programa permite al usuario elegir entre reconocer el texto completo o el área seleccionada manualmente para cada página individual, o realizar un análisis de diseño de página en cada página para detectar automáticamente áreas de reconocimiento apropiadas.



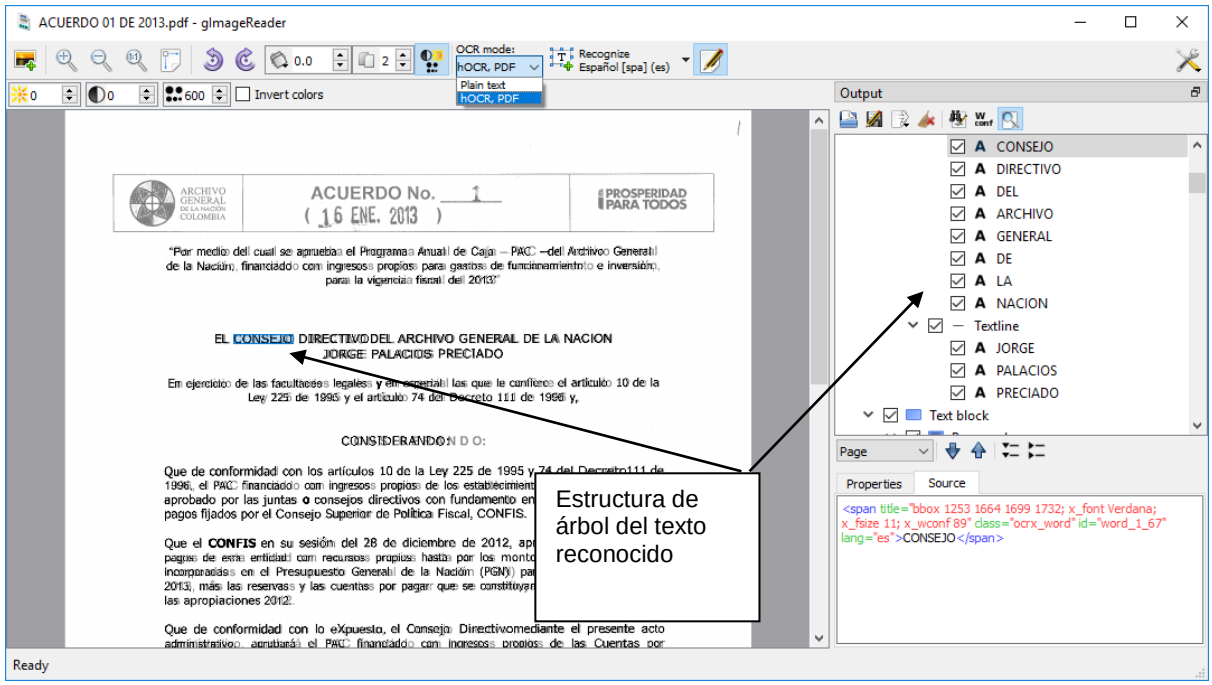
- El texto reconocido aparecerá en el panel de salida (a menos que el texto haya sido redirigido al portapapeles), que se muestra automáticamente en cuanto se reconoce algún texto.



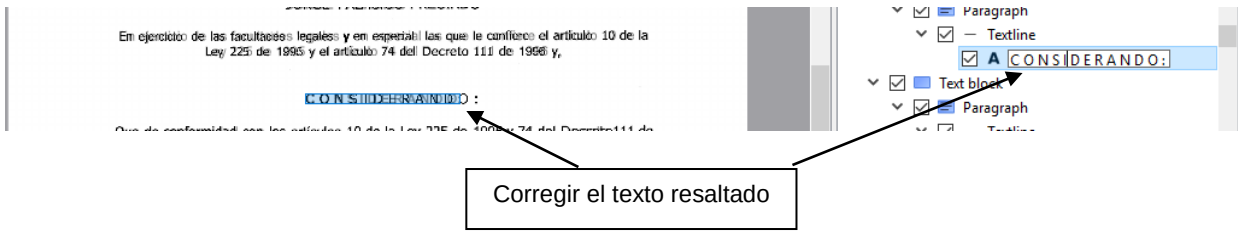
- Cuando se reconoce texto adicional, o bien se adjunta, se inserta en la posición del cursor o se reemplaza el contenido anterior del búfer de texto, dependiendo del modo seleccionado en el menú del modo de adición, que se puede encontrar en la barra de herramientas del panel de salida.

Reconocimiento y post-procesamiento en [i]hOCR, modo PDF

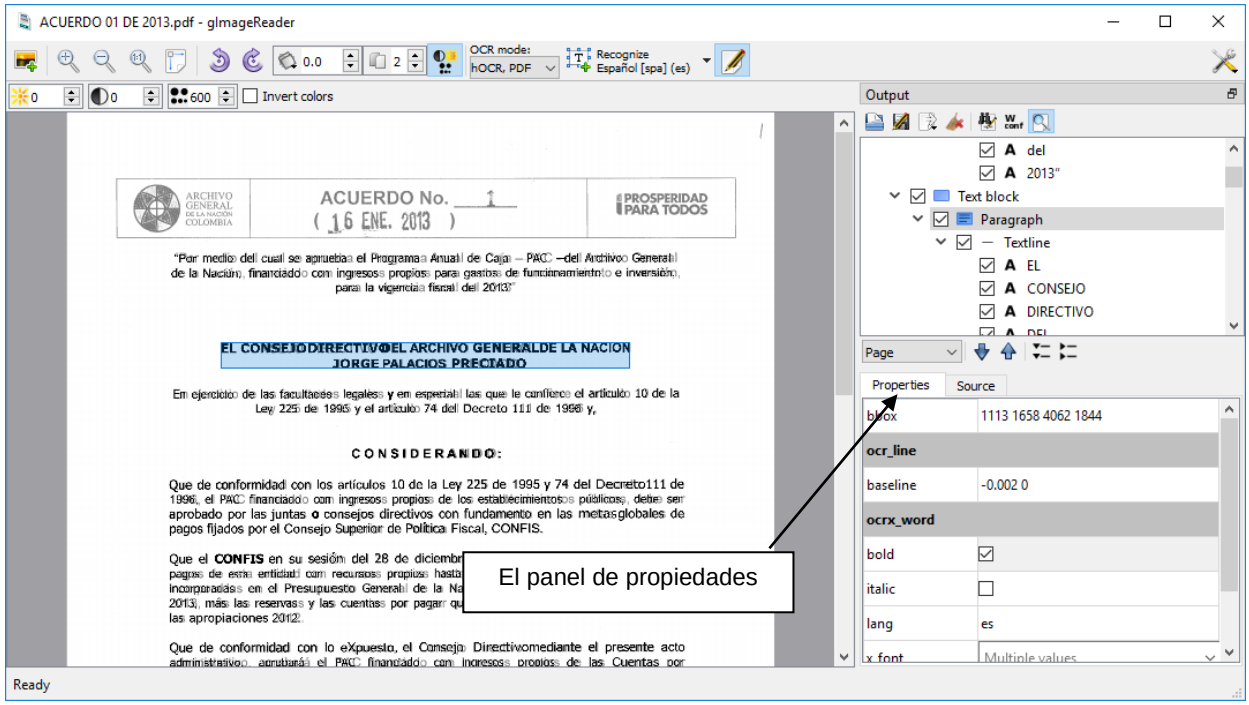
- En el modo hOCR, siempre se reconoce la página completa de la(s) fuente(s) seleccionada(s).
- El resultado de reconocimiento se presenta en el panel de salida como una estructura de árbol, dividida en páginas, párrafos, líneas de texto, palabras y gráficos.



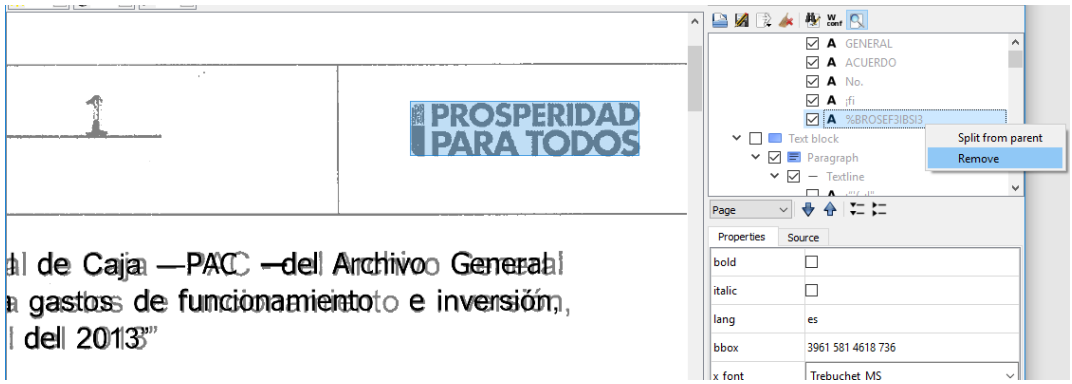
- Cuando se selecciona una entrada en la estructura de árbol, el área correspondiente se resalta en la imagen. Además, las propiedades de formato y diseño de la entrada se muestran en la pestaña Propiedades debajo del árbol del documento. La fuente de hOCR sin formato es visible en la pestaña Origen debajo del árbol del documento.
- El texto de la palabra en el árbol del documento se puede editar haciendo doble clic en la entrada de la palabra correspondiente. Si una palabra está mal escrita, se volverá roja. Al hacer clic con el botón derecho en una palabra en el árbol del documento se mostrará un menú con sugerencias de ortografía.



- Las propiedades de una entrada seleccionada se pueden modificar haciendo doble clic en el valor de la propiedad deseada en la pestaña Propiedades. Las acciones interesantes para las entradas de texto son ajustar el área delimitada, cambiar el idioma y modificar el tamaño de la fuente. La propiedad de idioma también define el idioma de ortografía utilizado para verificar la palabra respectiva. El área delimitadora también se puede editar cambiando el tamaño del rectángulo de selección en el lienzo.

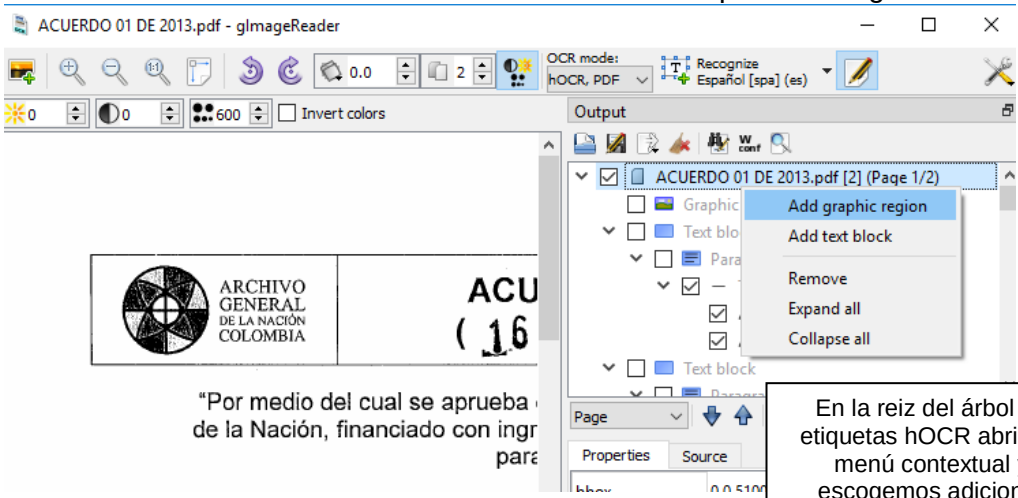


- Los elementos arbitrarios se pueden eliminar del documento haciendo clic derecho en el elemento correspondiente.



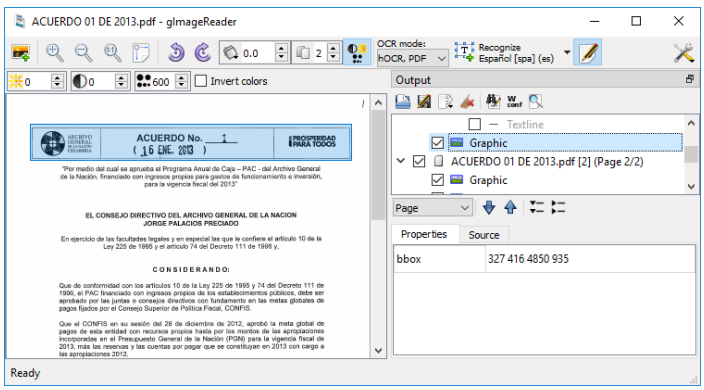
- Se pueden definir nuevas áreas gráficas seleccionando la entrada Agregar región gráfica del menú contextual del elemento de página respectivo y dibujando un rectángulo en el lienzo.

Esta funcionalidad se usa cuando el reconocimiento de texto se aplica a una gráfica o icono



En la reiz del árbol de etiquetas hOCR abrimos menú contextual y escogemos adicionar región grafica

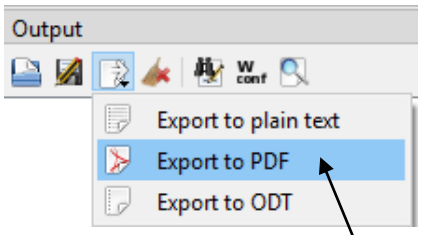
El resultado se ve reflejado al exportar el documento a pdf, en donde se exportan los gráficos tal cual los tenía la imagen original



- El árbol de documentos se puede guardar como un documento hOCR HTML a través del botón Guardar como hOCR en la barra de herramientas del panel de salida. Los documentos existentes pueden importarse a través del botón Abrir archivo hOCR en la barra de herramientas del panel de salida.

Botón guardar como hOCR

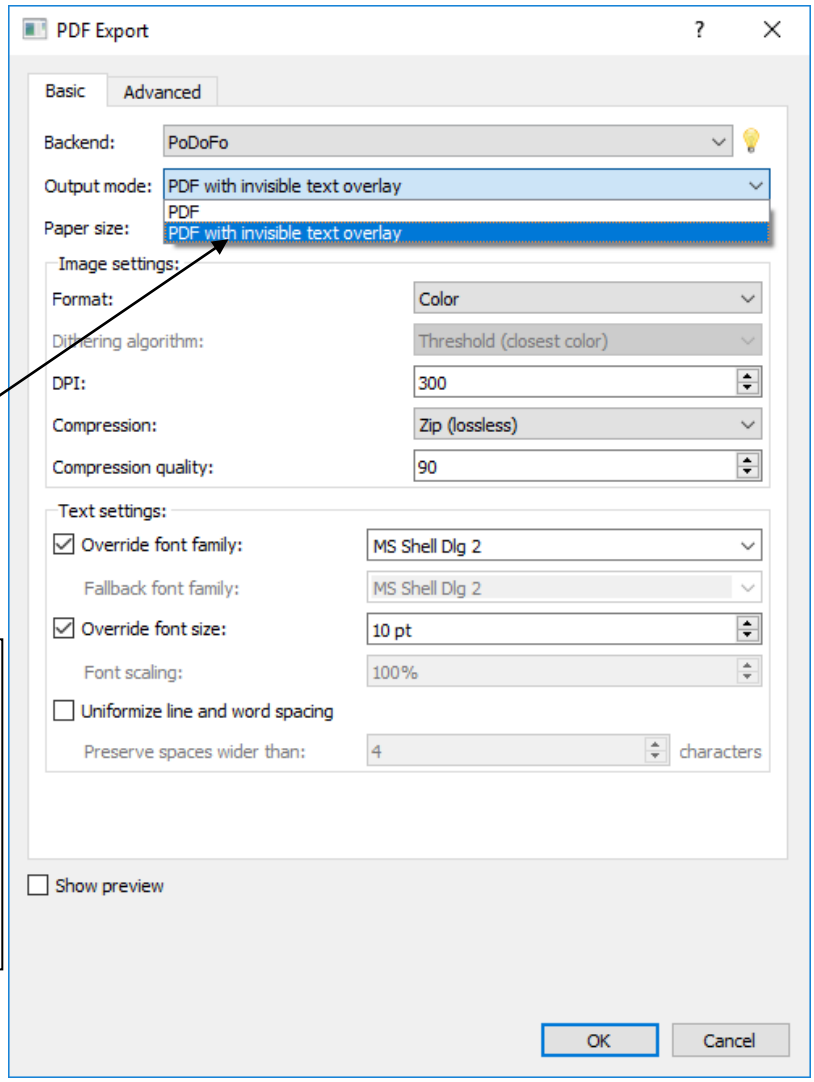
- Los archivos PDF se pueden generar desde el menú de exportación PDF en la barra de herramientas del panel de salida. Dos modos están disponibles:
    - PDF generará un PDF reconstruido con el mismo diseño y gráficos / imágenes que el documento fuente.
    - El PDF con superposición de texto invisible generará un PDF con la imagen original no modificada como fondo y texto invisible (pero seleccionable) superpuesto encima del texto fuente respectivo en la imagen. Este modo de exportación es útil para generar un documento que es visualmente idéntico a la entrada, pero con texto que se puede buscar y seleccionar.



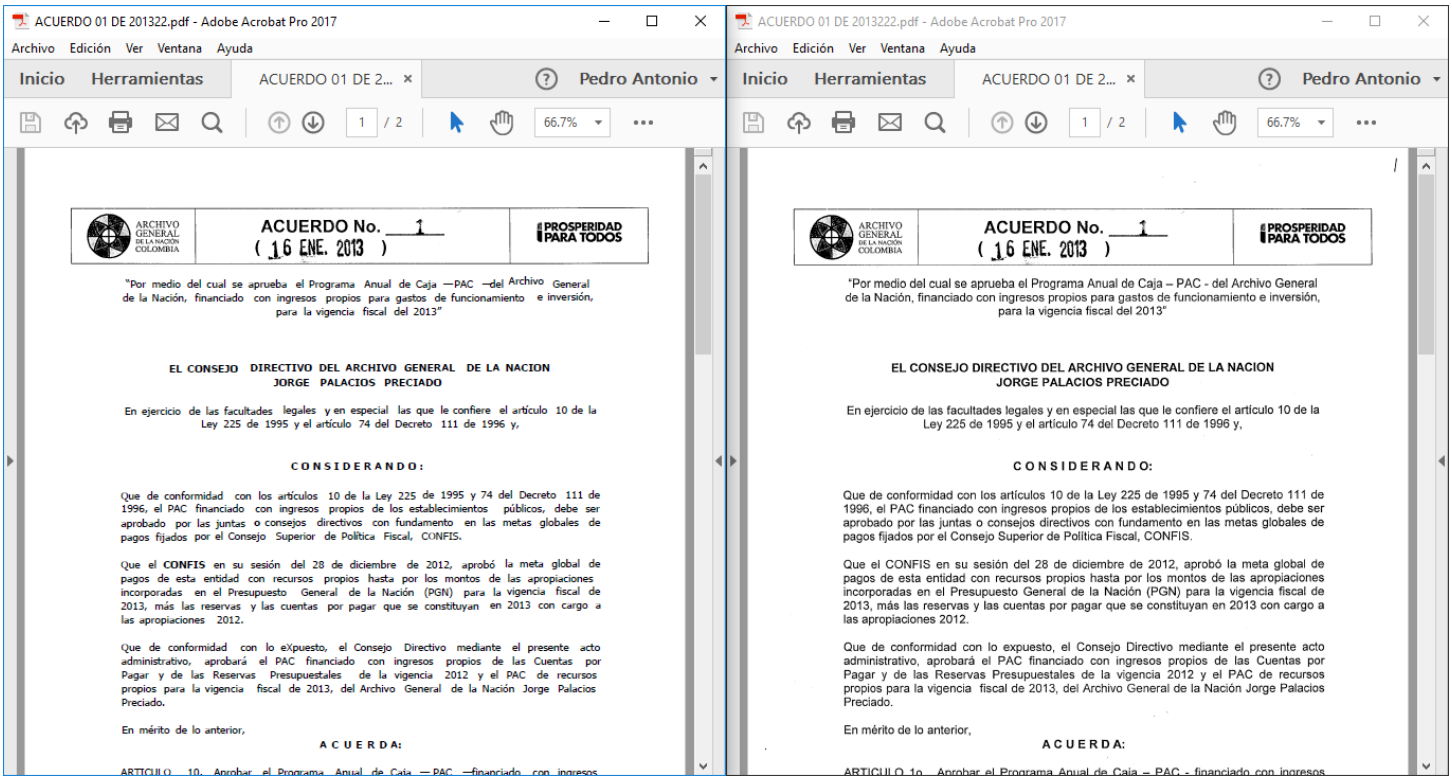
Botón exportar a PDF

El modo de salida permite escoger entre generar un PDF nativo con texto limpio, o generar el mismo PDF de origen con el texto reconocido encima del texto original en modo invisible.

Al exportar a PDF, se le solicita al usuario que use la familia de fuentes, si respeta los tamaños de fuente detectados por el motor de OCR y si intenta homogeneizar el espaciado entre líneas de texto. Además, el usuario puede seleccionar el formato de color, la resolución y el método de compresión para usar en las imágenes del documento PDF para controlar el tamaño de la salida generada.



Aquí se muestran los dos resultados a la izquierda se muestra el texto limpio reconocido y a la derecha el documento original con el texto reconocido invisible encima del texto original.



[i] hOCR es un estándar abierto de representación de datos para texto con formato obtenido del reconocimiento óptico de caracteres (OCR).