3C Tecnología *Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33*
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

Recepción: 24 de octubre de 2013
Aceptación: 27 de diciembre de 2013
Publicación: 25 de febrero de 2014

DOCLUX OCR: SOFTWARE LIBRE PARA LA RESTAURACIÓN Y TRANSCRIPCIÓN DE IMÁGENES ARCHIVÍSTICAS

DOCLUX OCR: FREE SOFTWARE FOR IMAGE RESTORATION AND TRANSCRIPT ARCHIVISTIC

Lisbet Milagros Guerra Cantero [1]

Eriberto Vanegas Lago [2]

1. Ingeniera en Ciencias Informáticas. Profesor del Departamento de Web y Multimedia. Facultad Regional de Ciencias Informáticas de la Universidad de Granma. Cuba. E-mail: lmcantero@grm.uci.cu
2. Ingeniero en Ciencias Informáticas. Profesor del Departamento de Soluciones de Gestión. Facultad Regional de Ciencias Informáticas de la Universidad de Granma. Cuba. E-mail: evanegas@grm.uci.cu

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

RESUMEN

El Archivo Histórico del Municipio de Manzanillo se atesora gran cantidad de documentos, esta institución tiene la misión de recuperar, restaurar y difundir la documentación de valor histórico que conserva. Cuenta con un portal web a disposición del público, donde se pueden consultar imágenes digitalizadas de dicha información. Sin embargo, producto al grado de deterioro que presentan dichos documentos antiguos, es necesario realizarles un proceso de restauración para mejorar su calidad. Producto a esto se desarrolló el sistema DocLux que permite el tratamiento de imágenes digitalizadas por lotes. DocLux permite aplicar una serie de filtros que posibilitan mejorar la calidad de las imágenes. A pesar de esto, es necesario aplicarle un proceso de transcripción, ya que se dificulta el reconocimiento de la información contenida en las imágenes. Surge la necesidad de desarrollar una alternativa de software libre que permita el Reconocimiento Óptico de los Caracteres. Se desarrolló un prototipo de aplicación DocLux OCR que permite reconocer y transcribir las vocales minúsculas sin tilde del lenguaje español, utilizando el motor Tesseract.

ABSTRACT

The Historical Archive of the Municipality of Manzanillo is treasured lot of documents, this institution has the mission to recover, restore and dissemination of records of historical value it retains. It has a website available to the public, where you can view digitized images of such information. However, due to the degree of impairment arising from those old documents, you need to have a follow restoration process to improve its quality. Product of this system was developed which allows the treatment DocLux digitized image batch. DocLux allows applying a series of filters which allow improving the quality of images. Despite this, it is necessary to apply a process of transcription, as it makes the recognition of the information contained in the images. The need arises to develop a free software alternative that allows the Optical Character Recognition. We developed a prototype application that DocLux OCR to recognize and transcribe the vowels tiny Spanish language without accent, using the Tesseract engine.

PALABRAS CLAVE

Restauración, tratamiento, transcripción.

KEY WORDS

Restoration, processing, transcription.

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

INTRODUCCIÓN

Actualmente es apremiante la digitalización de los documentos custodiados en los archivos, así como la recuperación y restauración de la información de valor histórico que en ellos se conservan. Debido a la antigüedad de los documentos, la manipulación de los investigadores, la humedad, las manchas y los huecos, esta valiosa fuente de conocimiento podría desaparecer.

ARCHIVO HISTÓRICO DE MANZANILLO

En el Archivo Histórico del Municipio de Manzanillo (AHMM), se atesora gran cantidad de documentos desde finales de los siglos 70. Esta institución tiene la misión de recuperar, restaurar y difundir la documentación de valor histórico que en ella se conserva, así como ponerla a disposición del público. Actualmente cuenta con un portal web, posibilita consultar las imágenes digitalizadas de los documentos históricos. Sin embargo, producto al grado de deterioro (los efectos de envejecimiento, la filtración de la tinta, las manchas, los huecos, la suciedad, la decoloración del papel) que presentan los documentos, se hace necesario realizar un proceso de restauración, que consiste en mejorar su visualización y comprensión a través de las herramientas Gimp y Photoshop. Este procedimiento demora y dificulta el proceso debido al enorme volumen de información existente. Además, generalmente las imágenes no terminan con la calidad requerida pues se necesita habilidad y conocimiento de dicha herramienta para llevar a cabo esta tarea.

Como solución a este problema se desarrolló en el Centro de Desarrollo de la Facultad Regional de Ciencias Informáticas de la Universidad de Granma, el sistema DocLux que permite el tratamiento de imágenes digitalizadas por lotes. Este sistema permite cargar una o varias imágenes a las cuales, para mejorar su visualización, se les pueden aplicar una serie de filtros: brillo, contraste, balancear colores, modificar colores, escala de grises; algunos de estos de forma manual y/o automática. Además incorpora las herramientas rotar y escalar la imagen. Las imágenes obtenidas finalmente pueden ser almacenadas a elección del usuario en formato PDF o en cualquiera de los siguientes formatos: jpg, jpeg, png, bmp, y tiff. Cada una de las imágenes procesadas por DocLux contendrá una marca de agua que identifica al AHMM y que garantiza que no ha sido modificada (Ibar, Correoso 2013).

Luego de aplicarle el proceso de tratamiento a las imágenes digitalizadas se evidencia una considerable mejora en la calidad de las mismas. Sin embargo, estos documentos antiguos presentan en su gran mayoría tipografías extrañas y sintaxis obsoletas, lo cual dificulta una adecuada compresión del texto contenido en la imagen.

Teniendo en cuenta los problemas identificados surge la necesidad de desarrollar un sistema con herramientas libres que permita transcribir documentos antiguos y manuscritos utilizando Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés).

RECUPERACIÓN Y RESTAURACIÓN DE IMÁGENES

Luego de digitalizar los documentos a través de una cámara fotográfica o escáner es necesario realizar el proceso de restauración. Se puede considerar que este proceso se realiza con el objetivo de reconstruir o recuperar una imagen que ha sido distorsionada o degradada.

Los archivos históricos conservan, por lo general, una gran cantidad de documentos. El proceso de restaurar las imágenes de no realizarse de forma automática puede consumir recursos materiales y humanos. Una de las tendencias más utilizadas es la integración de sistemas de restauración de imágenes con motores OCR. Estos últimos permiten capturar e identificar texto incluido dentro de las imágenes, puede definirse como el método o técnica para reconocer la parte textual de un documento digitalizado (Charca 2007).

El proceso OCR permite editar el texto que generalmente se encuentra bloqueado dentro de las imágenes digitalizadas. Funciona utilizando un tipo de inteligencia artificial conocido como reconocimiento de formas y estructuras, que identifica los caracteres individuales de texto en el documento incluyendo los espacios en blanco, signos de puntuación y finales de líneas. Este proceso recibe como entrada un documento digitalizado y obtiene como salida un archivo de texto que puede ser editables. El proceso PCR se basa en 5 etapas: adquisición de la imagen, binarización, fragmentación o segmentación de la imagen, representación digital o extracción de características y distinción del carácter contenido en la imagen o reconocimiento (Fernández, Consuegra 2008).

Con la aparición de los algoritmos de OCR han surgido varios software que complementan el proceso de la restauración. En el siguiente epígrafe se tratan sistemas estudiados referentes al reconocimiento de caracteres.

3C Tecnología *Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33*
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

SOLUCIONES EXISTENTES PARA RESTAURACIÓN DE IMÁGENES

ABBYFineReader es un programa de OCR, capaz de reconocer 176 lenguajes, incluso combinados en una misma página. Recibe como formato de entrada documentos impresos en digitales y se obtiene como salida formatos como: PDF, HTML, RTF, TXT, entre otros. Convierte los documentos escaneados o los PDF en texto editable, es decir, extrae el texto de los documentos digitalizados y crea archivos PDF de búsqueda para archivar. El mismo tiene como requisitos de sistema 110MB de espacio libre en disco, sistema operativo Windows 95 (OSR2 with USB- support) /98/ME/2000, Kofax Ascent Capture 5.0 instalado y se recomienda: Intel Pentium II o mayor, Windows 2000, 128MB de memoria. Además tiene un costo de $129 (González, Villaverde, Mesa 1996).

Omnipage professsional 11 para Windows Castellano es un programa para escáner fabricado por Scansoft que reconoce 100 idiomas incluso cuando se tienen varios en una misma página. Recibe como formato de entrada PDF, documentos impresos digitales, TIFF-FX y obtiene como salida los formatos: PDF, HTML, escanea directamente a cualquier editor de texto compatible con Microsoft Office 2000. El mismo tiene como requisitos de sistema disco duro: 140MB, memoria RAM: 32 MB (se recomiendan 64 MB), procesador: Intel Pentium o equivalente, sistema operativo: Windows 95, 98, 2000, ME y NT 4.0 o XP, tarjeta gráfica: Monitor SVGA, 800 x 600 como mínimo (256 colores o más) y hardware adecuado para la captura de imágenes. Además un costo de $539,21 (Gómez 2004).

En la web también aparecen varios sitios de OCR Online que permiten cargar imágenes y extraer el texto editable de forma muy explícita e intuitiva.

Free Online OCR es un sistema que permite reconocer texto y caracteres de los documentos PDF, fotografías y cámaras digitales. Es un servicio gratuito en un “modo invitado” que soporta 32 idiomas de reconocimiento. Extrae el texto de imágenes con formatos JPG, JPEG, BMP, TIFF, GIF y obtiene como salida el texto en Microsoft Word, Microsoft Excel y TXT.

Softi FreeOCR es un programa que sirve para interpretar documentos escaneados y convertirlos a texto editable. Si se cuenta con documentos en papel y se necesita digitalizarlo resulta una solución muy efectiva y práctica, permitiendo ahorrar un tiempo considerable. Sin embargo el motor que utiliza sólo incluye por defecto el reconocimiento en inglés, aunque se puede añadir paquetes para el español.

FreeOCR es un sistema totalmente gratuito. Es compatible con los formatos de imagen JPG, GIF, TIFF y BMP, así como PDF y texto multi-columna. Puede reconocer hasta 30 lenguajes. Como limitación tiene que no permite imágenes de peso superior a 2MB y sólo permite subir 10 imágenes por hora.

Online-OCR es un programa que permite convertir fotografía e imágenes de captura con cámaras digitales en texto con 32 lenguajes de reconocimiento. Extrae texto de imágenes con formato JPG, JPEG, BMP, TIFF y GIF y se obtiene como resultado archivos editables Word, EXCEL, TXT, PDF, HTML. Como limitación tiene que sólo puede subir 15 imágenes por hora.

Aunque las herramientas de OCR antes mencionadas posibilitan un adecuado rendimiento en cuanto a documentos impresos modernos, no se puede dejar de mencionar que sus tasas de errores son elevadas cuando se trata de tipografía antigua o documentos manuscritos, lo cual los descarta como una opción. A pesar de que el texto se compone básicamente de caracteres individuales, la mayoría de los algoritmos de OCR no consiguen buenos resultados en contextos continuos o manuscrito. Esto se debe a que los documentos antiguos presentan en su mayoría tipografías extrañas, sintaxis obsoletas, además de verse afectados por humedad, manchas, agujeros que dificultan una correcta transcripción. Por lo que no se puede hablar de un sistema convencional, sino de un sistema que recogería además otras funcionalidades que permitan una transcripción adecuada para el tipo de documento al que se refiere.

4State: Sistema Multimodal de transcripción asistida de documentos antiguos es el resultado de la cooperación entre el grupo de investigación en "Percepción y aprendizaje computacionales" de la Universidad de Jaume I y 4TIC. El sistema supera las dificultades anteriormente mencionadas de la siguiente forma (4TIC 2009).

Frente a posibles fuentes de ruido (humedad, manchas, agujeros), integra una serie de herramientas de tratamiento y análisis de la imagen que facilita la preparación de las páginas antes del proceso de OCR. Dicha preparación puede incluye tanto la limpieza de la imagen, como la detección del diseño o estructura del texto, en particular su división en líneas. Frente a las dificultades propias de la variabilidad de la caligrafía del texto manuscrito, así como las peculiaridades tipográficas, léxicas y sintácticas, el sistema ofrece una variedad de reconocedores OCR específica y automáticamente adaptable a las características de cada tarea de transcripción a la que haya que enfrentarse. Frente a la inevitable aparición de errores en el resultado del proceso de OCR, el sistema permite un entorno gráfico pensado para facilitar al máximo la supervisión humana de las transcripciones automáticas y la eventual corrección de los errores cometidos (4TIC 2009).

En el funcionamiento del sistema intervienen dos aplicaciones que pueden ejecutarse simultáneamente en máquinas distintas. StateTA, es una aplicación interactiva, controlada mediante un lápiz electrónico y una pantalla táctil para ayudar a los usuarios en la transcripción de documentos antiguos. Todas las transcripciones obtenidas como resultado se firman digitalmente junto a la original, para garantizar que no se modifiquen en el tiempo (4TIC 2009).

Luego de analizar las características generales de las diferentes herramientas estudiadas se determinó que estas soluciones informáticas son privativas y no están al alcance de todos los necesitados y las libres encontradas no solucionan el problema existente en el AHMM. Por lo que esta investigación se enfoca en las alternativas de software libre que permitan resolver el problema de la restauración de las imágenes archivísticas de documentos antiguos y manuscritos. Por lo antes expuesto se llega a la conclusión de la necesidad de desarrollar un sistema libre para la restauración de imágenes archivísticas que responda a los problemas actuales del AHMM.

HERRAMIENTAS Y METODOLOGÍA

Para dar solución a las necesidades antes mencionadas se realizó un estudio de las herramientas libres que se utilizarán para desarrollar el sistema DocLux.

MOTOR OCR

Muchas de las aplicaciones existentes en el mundo hacen uso de motores OCR por la precisión y seguridad con la que trabajan. Para el desarrollo del sistema se realizó un estudio del cual se seleccionó el motor OCR Tesseract. El mismo tiene implementada una red neuronal, que permite el reconocimiento acertado de caracteres. Es capaz de leer una amplia variedad de formatos de imagen y convertirlos a texto en más de 40 idiomas.

LENGUAJE DE PROGRAMACIÓN

Luego del estudio realizado se identificó que los lenguajes C, C++, Java y Python, son los más utilizados en el desarrollo de aplicaciones para el tratamiento de imágenes. Finalmente se decide utilizar C++, principalmente porque es el lenguaje en el que está implementado el motor OCR a utilizar. Además por la diversidad de bibliotecas para el tratamiento de imágenes con las que cuenta, tales como: PNGwriter, FreeImage, ImgSource, Magick++, Cimg, entre otras que fueron utilizadas en el sistema a desarrollar.

FRAMEWORK PARA EL DESARROLLO DE APLICACIONES

Para el desarrollo de la aplicación se utilizó el framework Qt, el cual está diseñado para ser ejecutado en una variedad de sistemas operativos. Utiliza C++ como lenguaje nativo, aunque ofrece soporte para otros como Python, Java o C# mediante PyQt, QtJambiQyoto respectivamente.

ENTORNO DE DESARROLLO INTEGRADO (IDE)

Como IDE se utilizó Qt Creator pues es multiplataforma y se utiliza a las necesidades del cliente. Además cuenta con un editor de código con soporte para C++, auto-completado de código, herramientas para la rápida navegación del código, resaltado de sintaxis, permite realizar programación visual y dirigida por eventos, así como un elevado tiempo de respuesta y no ocupa gran cantidad de memoria.

METODOLOGÍA

Para guiar y facilitar el trabajo de los desarrolladores en la creación y documentación del software se utilizó la metodología híbrida cubana SXP. Está compuesta por la metodología SCRUM y XP. Esta ofrece una estrategia tecnológica, a partir de la introducción de procedimientos ágiles que permitan actualizar los procesos de software para el

mejoramiento de la actividad productiva. SXP agrupa las mejores prácticas de ambas metodologías (Abad, Romero, Villar, Céspedes, García 2009).

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

RESULTADOS Y DISCUSIÓN

Una vez aplicado el proceso de tratamiento a las imágenes digitalizadas, que permite mejorar considerablemente la calidad y visualización de las mismas, se hace necesario aplicarle el proceso de transcripción ya que a pesar de la mejora apreciada en la imagen, se dificulta el reconocimiento de la información contenida en las imágenes, además para cumplir con la misión de los archivos históricos de preservar y difundir la memoria documental que custodian.

Como solución a este problema identificado se decidió desarrollar un prototipo de aplicación como paso inicial para el reconocimiento de los documentos antiguos y manuscritos custodiados en esta institución. Como base el prototipo utiliza el motor Tesseract, el cual permite reconocer y transcribir las vocales minúsculas sin tilde del lenguaje español.

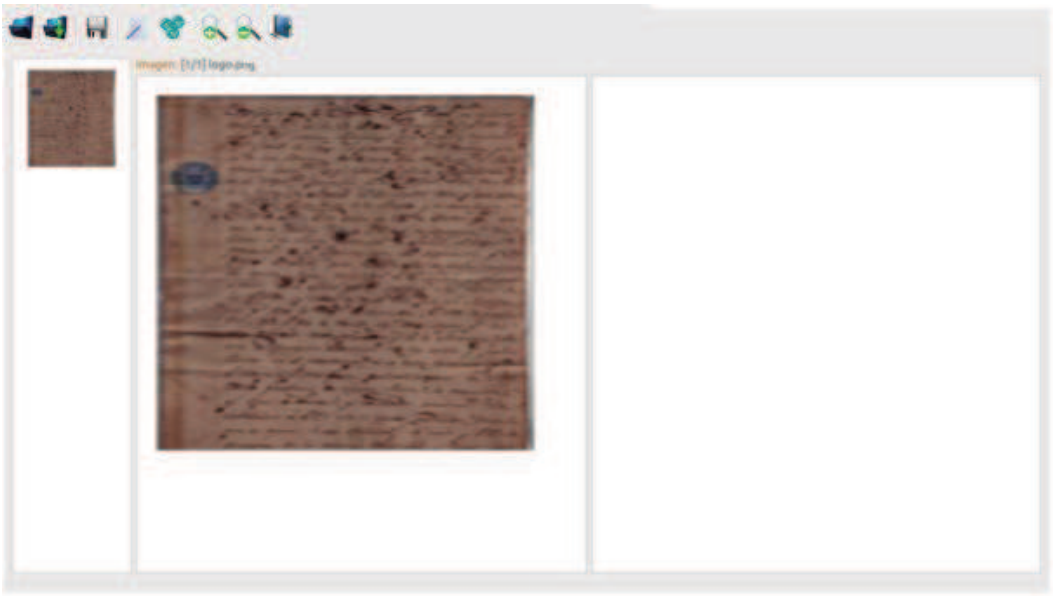

Gráfico 1. Interfaz gráfica de DocLux OCR. **Fuente:** Elaboración propia.

El funcionamiento básico del prototipo consiste en un visor de imagen donde el usuario podrá visualizar la imagen actual a procesar y un visor de texto donde se visualizará el texto resultante luego de ser aplicado el procedimiento de OCR. Contiene una barra de herramientas con cada una de las funcionalidades que permitirá la aplicación (Villariño, Vallés 2013).

3C Tecnología Vol.3 – Nº 1 | Febrero – mayo 2014, 21 - 33
Área de Innovación y Desarrollo, S.L.
ISSN: 2254 – 4143

Gráfico 2. Delimitación de las áreas de DocLux OCR**. Fuente:** Elaboración propia.

La aplicación cuenta con los siguientes requisitos funcionales:

- **Cargar imagen:** permite cargar una o varias imágenes a la aplicación con cualquiera de los siguientes formatos: JPEG, JPG, PNG, BMP, TIFF, PPM, XBM o XPM.
- **Agregar imagen:** permite agregar una o varias imágenes a la lista de imágenes cargadas.
- **Ampliar/Reducir tamaño:** permite ampliar/reducir el tamaño de la imagen actual que se procesa.
- **Aplicar OCR:** permite aplicarle el procedimiento OCR a la imagen actual.
- **Guardar archivo OCR:** permite guardar en la dirección especificada por el usuario el texto resultante del procedimiento OCR en un archivo de texto con cualquiera de los siguientes formatos: ODT, DOC o TXT.
- **Cargar archivo:** permite cargar un archivo de texto a la aplicación con cualquiera de los siguientes formatos: ODT, DOC o TXT.
- **Editar texto resultante:** permite editar el archivo resultante o el archivo cargado.

Como requerimientos no funcionales de hardware la aplicación debe contar con un microprocesador Pentium IV de 1,6 GHz, memoria RAM de 256 MB o superior, disco duro de 20 GB o superior. Además, como sistema operativo GNU/Linux, debe tener instalado Qt Creator en su versión 2.4.1, la biblioteca del motor OCR Tesseract en su versión 3.02 y el

framework Qt en su versión 4.8.0. El tamaño de la aplicación es de 2MB, lo que permite ser portado en cualquier dispositivo y no requiere de gran espacio para ser instalado.

El prototipo de aplicación desarrollado como paso inicial para la realización de un sistema con herramientas libre para el AHMM cumple con las metas propuestas, pues permite reconocer las vocales minúsculas de los documentos antiguos o manuscrito existentes en el archivo histórico, lo que representa un importante paso, para conservar y difundir la documentación que en ellos se preserva.

CONCLUSIONES

Durante el desarrollo de la presente investigación se puso de manifiesto la necesidad de desarrollar un sistema informático con herramientas libres que permita transcribir documentos antiguos y manuscritos utilizando el OCR.

- Se demostró que existen varios sistemas libres para la transcripción de documentos digitalizados, pero estos no satisfacen las necesidades del AHMM, ya que no realizan una adecuada transcripción de los documentos antiguos y manuscritos.
- Se identificaron otras herramientas que cumplen con varias de las necesidades actuales del AHMM, sin embargo las mismas son privativas y no están al alcance de los archivos históricos de países pobres, como lo es Cuba.
- Con el desarrollo del sistema para la restauración de imágenes archivísticas, utilizando software libre, se logra dar un paso significativo para resolver el problema existente en el AHMM permitiendo reconocer las vocales minúsculas sin tilde del lenguaje español.
- Mediante el estudio de los sistemas analizados como 4State se pudieron identificar nuevas funcionalidades que en versiones futuras pudiera contener el sistema desarrollado.

REFERENCIAS

- 4TIC. *4State: Sistema Multimodal de transcripción asistida de documentos antiguos* [en línea]. Lugar: Universidad de Jaume I, 2009. [Consulta: 15 de Noviembre de 2012]. Disponible en: <www.4tic.com>
- Abad Meneses, Abel; Marsi Romero, Gladys Peñalver, Villar, Malay Rodríguez, Céspedes, Raycel Fernández and García, Susel Pino. *SXP. Metodología ágil para proyectos de software libre*. *Unicornios*, 38, 2009, p. 17-30.
- Ayala Charca, Giancarlo. *Reconocedor óptico de caracteres de placas de automóviles empleando técnicas de procesamiento digital de imágenes*. Universidad Nacional de Ingeniería, 232, 2007, p. 64-97.
- Sánchez Fernández, Carlos Javier y Sandonís Consuegra, Víctor, 2008. *Reconocimiento Óptico de Caracteres (OCR).* Universidad Carlo III, 7, 2008, p. 2-5
- Flores Gómez, Miguel Eduardo. *Reconocedor Óptico de Caracteres Manuscritos (ROCM) por medio de redes neuronales*. Universidad Don Bosco, 176, 2004, p. 88-91.
- Hilera González, José R; Romero Villaverde, Juan P y de Gutiérrez Mesa, José A. *Sistema de Reconocimiento Óptico de Caracteres (OCR) con Redes Neuronales.* 1996.
- Nariño Ibar, Thaylin y González Correoso, Uberlandys. *Tratamiento de imágenes de archivos por lotes (v2.0) DocLux*. Universidad de las Ciencias Informáticas, 136, 2013, p. 64-97.
- Mena Mugica, Mayra. *Gestión Documental y organización de archivos*. Lugar: Habana, Cuba. Editorial: Félix Varela, 2005. ISBN 959-258-950-X.
- Galán Villariño, Yilian y Muñoz Valles, Gerardo José. *Prototipo de aplicación para reconocer y transcribir las vocales minúsculas sin tilde del lenguaje español para DocLux*. Universidad de las Ciencias Informáticas, 99, 2013, p. 6-55.