Nuevos enfoques del OCR para los libros impresos antiguos

New Approaches to OCR for Early Printed Books

Nikolaus Weichselbaumer, y otros

DigItalia, ISSN 1972-6201, Vol. 2, 2020, p. 74-87

Se presenta el proyecto OCR-D, financiado por la Deutsche Forschungsgemeinschaft (DFG) y puesto en marcha en 2014, a partir de la organización de un taller en el que expertos historiadores del libro e informáticos evaluaron los nuevos avances en el reconocimiento óptico de caracteres (OCR). En las últimas décadas, muchas bibliotecas han empezado a digitalizar sus fondos de impresos antiguos, sin embargo, quedaba por desarrollar la estrategia tecnológica que permitiera que los documentos digitalizados estén disponibles para la búsqueda a texto completo y su posterior procesamiento con herramientas de las Humanidades Digitales. El reconocimiento del texto completo de los documentos históricos es especialmente complicado debido a su gran variabilidad en cuento a letra, diseño, idioma y ortografía. Además, los motores OCR suelen estar entrenados con los tipos de letra actuales por lo que se ignora la gran variedad regional y estilística de la tipografía de los impresos anteriores a 1800. El proyecto OCR-D, llevado a cabo por la Academia de Ciencias y Humanidades de Berlín-Brandeburgo, la Biblioteca Herzog-August de Wolfenbüttel, la Biblioteca Estatal de Berlín y el Instituto Tecnológico de Karlsruhe, tiene como objetivo la creación de un marco conceptual y técnico que permita la transformación del texto completo de cualquier copia digital. El proyecto se organiza por fases, siendo la primera de ellas la creación de una herramienta que identifique automáticamente grupos de fuentes en imágenes de documentos antiguos, centrados en los grupos de fuentes góticas que se utilizaban habitualmente en los textos alemanes impresos en los siglos XV y XVI: Fraktur, Bastarda, Rotunda, Textura y Schwabacher. La herramienta fue entrenada con 35.000 imágenes y alcanza un nivel de precisión del 98%. No sólo puede diferenciar entre los grupos de fuentes antes mencionados, sino también entre las hebreas, griegas, anticuadas e itálicas. Desde que las tecnologías OCR empezaron a adoptar las redes neuronales profundas, existen varios motores OCR de código abierto que pueden adaptarse a diferentes tipos de documentos. En este sentido se ha desarrollado la segunda fase del proyecto, creando “okralact”, una infraestructura que permite utilizar varios de estos motores OCR de software libre como Tesseract, OCRopus, Kraken y Calamari. Al mismo tiempo facilita el entrenamiento para modelos específicos de grupos de fuentes. Se trabaja, así mismo en la posibilidad de diferenciar la tipografía de los diferentes talleres impresores, lo cual podría aportar mucha luz sobre algunas lagunas en la investigación histórica. Para finaliza, se trabaja en la puesta a disposición del software, para el público de forma gratuita.

DigItalia, ISSN 1972-6201, Vol. 2, 2020, p. 74-87

Se presenta el proyecto OCR-D, financiado por la Deutsche Forschungsgemeinschaft (DFG) y puesto en marcha en 2014, a partir de la organización de un taller en el que expertos historiadores del libro e informáticos evaluaron los nuevos avances en el reconocimiento óptico de caracteres (OCR). En las últimas décadas, muchas bibliotecas han empezado a digitalizar sus fondos de impresos antiguos, sin embargo, quedaba por desarrollar la estrategia tecnológica que permitiera que los documentos digitalizados estén disponibles para la búsqueda a texto completo y su posterior procesamiento con herramientas de las Humanidades Digitales. El reconocimiento del texto completo de los documentos históricos es especialmente complicado debido a su gran variabilidad en cuento a letra, diseño, idioma y ortografía. Además, los motores OCR suelen estar entrenados con los tipos de letra actuales por lo que se ignora la gran variedad regional y estilística de la tipografía de los impresos anteriores a 1800. El proyecto OCR-D, llevado a cabo por la Academia de Ciencias y Humanidades de Berlín-Brandeburgo, la Biblioteca Herzog-August de Wolfenbüttel, la Biblioteca Estatal de Berlín y el Instituto Tecnológico de Karlsruhe, tiene como objetivo la creación de un marco conceptual y técnico que permita la transformación del texto completo de cualquier copia digital. El proyecto se organiza por fases, siendo la primera de ellas la creación de una herramienta que identifique automáticamente grupos de fuentes en imágenes de documentos antiguos, centrados en los grupos de fuentes góticas que se utilizaban habitualmente en los textos alemanes impresos en los siglos XV y XVI: Fraktur, Bastarda, Rotunda, Textura y Schwabacher. La herramienta fue entrenada con 35.000 imágenes y alcanza un nivel de precisión del 98%. No sólo puede diferenciar entre los grupos de fuentes antes mencionados, sino también entre las hebreas, griegas, anticuadas e itálicas. Desde que las tecnologías OCR empezaron a adoptar las redes neuronales profundas, existen varios motores OCR de código abierto que pueden adaptarse a diferentes tipos de documentos. En este sentido se ha desarrollado la segunda fase del proyecto, creando “okralact”, una infraestructura que permite utilizar varios de estos motores OCR de software libre como Tesseract, OCRopus, Kraken y Calamari. Al mismo tiempo facilita el entrenamiento para modelos específicos de grupos de fuentes. Se trabaja, así mismo en la posibilidad de diferenciar la tipografía de los diferentes talleres impresores, lo cual podría aportar mucha luz sobre algunas lagunas en la investigación histórica. Para finalizar, se trabaja en la puesta a disposición del software para el público de forma gratuita.

http://digitalia.sbn.it/article/view/2630

Resumen elaborado por María Osuna González

Libros de artista en exposición: recomendaciones de conservación preventiva

Andrea Paola Ruisanchez Campuzano

Ge-conservación, ISSN 1989-8568, Vol. 1, n. 18, 2020, p. 20-31

El presente artículo busca ser un referente para la conservación de un tipo de producciones artísticas con el que muchos conservadores no están familiarizados: los libros de artista. Estas obras plantean un reto de conservación que se acentúa en la toma de decisiones sobre cómo lograr su adecuada exposición sin comprometer su estabilidad material. Para el cabal disfrute y comprensión de dichos libros su manipulación es necesaria, haciendo que el conservador deba mediar entre el deterioro que esto conlleva y el respeto a su valor funcional, sus propiedades hápticas y la transmisión de su mensaje. La mediación también debe darse entre los usuarios y los custodios o propietarios de los libros, al igual que entre las instituciones que gestionan exposiciones. Esta investigación reflexiona sobre estos aspectos y busca plantear recomendaciones útiles para el conservador. Los libros de artista responden al libro objeto: piezas contemporáneas, de tirada corta, que adoptan la forma del libro como soporte y en el cual se plasman, a través de diferentes técnicas plásticas, conceptos visuales que se narran mediante una secuencia establecida, la cual comúnmente se compone de páginas. Para generar recomendaciones de conservación preventiva objetivas y eficientes se planteó la necesidad de evaluar los deterioros comunes causados a los libros de artista en exposición, y tratar de identificar los mecanismos y causas que los provocan; esto se logró mediante un muestreo y evaluación de una serie de ejemplares durante su manipulación en exposición. Se seleccionó un pequeño grupo de libros que funcionaran como ejemplo, para a través de la evaluación de su uso y alteraciones, poder establecer una relación con el resto de los libros de artista. La evaluación de los libros, su uso y las alteraciones causadas, se logró a través del diseño de una ficha clínica y una ficha de manipulación. Los campos que incluyó la ficha de manipulación fueron: facilidad de apertura del libro, nivel de manipulación del usuario, facilidad de interacción con ángulo de montaje, facilidad de manipulación con guantes de tela o látex, uso o interés del material de contexto (cédulas/explicaciones), confianza durante la manipulación del libro, y tiempo utilizado para la manipulación. Con el análisis comparativo de la interacción de los usuarios y los resultados de las fichas, se lograron agrupar en tres rubros las causas de deterioro y alteración que sufren este tipo de objetos: la manufactura del libro, en la que pueden englobarse dos asuntos: los materiales utilizados y la técnica de manufactura, esto

incluye deterioros inherentes a la pieza y sus materiales, que propicien la presencia de alteraciones sin necesidad de que el libro sea utilizado; montaje y museografía, ya que muchas veces el montaje de los libros durante la exposición puede ser un factor que propicie deterioros; y manipulación, porque los libros están concebidos y realizados por los artistas para ser observados y usados, pero cuando el número de asistentes en una exposición es elevado y una obra es manipulada constantemente durante numerosos días puede comprometerse su integridad. Es muy importante establecer que las recomendaciones de conservación deben ser modificadas y especificarse para cada colección, para cada exposición; esto será fundamentado, entre otras cosas, en el uso, valor y contexto del acervo de libros. El contenido a presentar son una serie de sugerencias puntuales que pueden realizarse antes de la manipulación de los usuarios, con el objetivo de dirigirla y encauzarla. Las recomendaciones de conservación serán presentadas en tres partes: recomendaciones para la preparación, deterioros detectados en los libros de artista evaluados después de su manipulación en exposición y exposición de libros de artista sin manipulación.

https://ge-iic.com/ojs/index.php/revista/article/view/689

Resumen elaborado por Antonio Rodríguez Vela

El proyecto inDICEs: medir el impacto de la cultura digital

Il progetto inDICEs: misurare l'impatto della cultura digitale

Sara Di Giorgio, Claudio Prandoni

DigItalia, ISSN 1972-6201, Vol. 2, 2020, p. 59-73p

Se presenta inDICEs como un proyecto de investigación e innovación, financiado por la Comisión Europea en el marco del programa Horizonte 2020, en respuesta a la convocatoria "Digitisation, Digital Single Market and European culture: new challenges for creativity, intellectual property rights and copyright". La Unión Europea considera el papel de las industrias culturales y creativas (ICC) como motores de la innovación económica y social. Los aspectos culturales y los sectores creativos contribuyen plenamente al desarrollo económico, generando empleo y crecimiento, y son por tanto cruciales para el futuro de Europa. En consecuencia, el sector del Patrimonio Cultural se considera un factor clave para el desarrollo de las ICC en Europa, por un lado, porque se proporciona acceso a grandes cantidades de contenidos reutilizables y por otro, porque se considera un laboratorio de

investigación y desarrollo del ecosistema cultural y creativo que contribuye al progreso económico y de la sociedad en general. En este sentido, inDICEs, nace con el objetivo de proporcionar a los gestores, tanto politicos como del sector ICC, herramientas para comprender plenamente el impacto social y económico de la digitalización del patrimonio cultural, además de evaluar la importancia de los contenidos abiertos para el desarrollo de nuevos modelos de negocio basados en la reutilización creativa, y analizar los problemas que surgen derivados de la gestión de derechos de autor. La investigación realizada por inDICEs, coordinada por el Istituto Centrale per il Catalogo Unico (ICCU) permitirá el desarrollo de una metodología científica para medir y evaluar el impacto económico de la digitalización del patrimonio cultural. El consorcio inDICEs reúne 14 organizaciones de 9 países europeos que conforman un conjunto multidisciplinar en el que se encuentran: institutos de investigación con presencia consolidada en los ámbitos del patrimonio cultural digital, las ciencias sociales y la propiedad intelectual; ONGs con capacidad de concienciación pública; representantes de las ICC y redes nacionales y paneuropeas; y empresas especializadas en el desarrollo de aplicaciones innovadoras, como Platoniq, una empresa con sede en Barcelona que desarrolla plataformas colaborativas para permitir la participación social y democrática de los ciudadanos y herramientas y metodologías digitales para la educación, la economía y la innovación social. InDICEs desarrollará además, una plataforma en línea que servirá de Observatorio Abierto, que proporcionará un conjunto de herramientas para elaborar estrategias de transformación digital de museos, bibliotecas y archivos.

http://digitalia.sbn.it/article/view/2628

Resumen elaborado por María Osuna González