Revista de Humanidades Digitales

READ: Recognition and Enrichment of Archival Documents

https://read.transkribus.eu/

Reseña realizada por:
Judit DE DIEGO MUÑOZ
Universidad Nacional de Educación a Distancia
juditdediego@gmail.com

1. INTRODUCCIÓN

La sinergia entre las nuevas tecnologías y las tareas puramente humanísticas o los objetos derivados de ellas son dos de los grandes intereses de las Humanidades Digitales, más específicamente, la ayuda que estas herramientas tecnológicas puedan prestar a la preservación del patrimonio histórico y cultural sea cual sea su soporte. La consecuencia más directa de esta simbiosis es la contribución a diferentes campos del conocimiento a través, por ejemplo, de la creación de herramientas en pro de la investigación. Y para muestra, un botón: Transkribus. El equipo de investigación responsable del proyecto READ[1] está asentado en Insbruck, Austria, y es la evolución de tranScriptorium[2]. READ crea en 2016 Transkribus, una plataforma de transcripción asistida pensada como herramienta de paleógrafos, archivistas, investigadores de Humanidades y Ciencias Sociales, así como el público general interesado en dichos campos del saber; de este último sector de público, especialmente, saldrán los voluntarios que ayuden al desarrollo del software.

Financiado por la Unión Europea y enmarcado en el programa de investigación e innovación Horizon 2020[3] (subvención número 674943), Transkribus nace de la necesidad de mejorar las técnicas de reconocimiento de texto manuscrito (HTR)[4] detectadas muchas ellas por los investigadores de READ.

[1] Por sus iniciales en inglés: Recognition and Enrichment of Archival Documents.
[2] Accesible desde: http://transcriptorium.eu/.
[3] Para saber más sobre este programa: https://ec.europa.eu/programmes/horizon2020/.
[4] Por sus iniciales en inglés: Handwritten Text Recognition.

ISSN 2531-1786

Revista Digital de Acceso Abierto http://revistas.uned.es/index.php/RHD/index
Recibido 07/04/2019 – Aceptado 10/07/2019

Así como de la aceptación del largo entrenamiento que aún le queda por recorrer al algoritmo para procesar, entender y decodificar correctamente el lenguaje humano debido a diferentes retos, como los estilos de escritura de cada autor o los caracteres propios de cada lengua. Sus metas, tal y como delata el nombre del proyecto, son el reconocimiento y el enriquecimiento del material de archivo.

2. ANÁLISIS DEL SITIO: ARQUITECTURA DE LA INFORMACIÓN, ACCESIBILIDAD Y EXPERIENCIA DE USUARIO

Read Revolutionizes Access to Handwritten Documents es el lema que abre la página web del proyecto READ, desarrollada sobre una sobria plantilla de fondo blanco que contrasta con pastillas más oscuras, negras o azul marino, y una tipografía de palo seco también de color negro: modelo típico de los sitios dedicados a proyectos oficiales.

Su arquitectura, basada en información fija, divide la información en ocho páginas principales bien comunicadas entre ellas gracias a enlaces internos que redireccionan, para ampliar la información, a la pestaña pertinente en cada caso. Estas son, como se ven en la figura 1: *Home, About, Network, News, Events,* Transkribus, *Publication* y *Contact*.

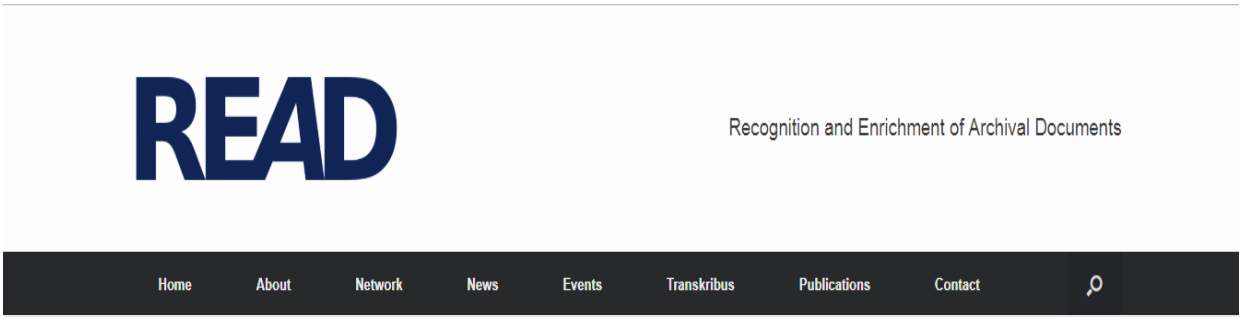

Figura 1. Cabecera de la web del proyecto *READ*.

Así, lo primero que encontramos es una página de inicio con una presentación icónica de los cuatro pilares en los que se sustentan la web y el propio proyecto: *About, Network, Research* y *Services*. Estas resumen la información total que se ofrece en la página, coincidiendo, solo en parte, con las pestañas del menú de navegación. Sin movernos de aquí hacemos *scroll* y encontramos un apartado final de texto líquido y actualizable, reservado para las últimas entradas relacionadas con el proyecto publicadas en la página *News*, correctamente enlazado con ella.

Antes de continuar navegando por el menú, se ha de advertir que, a partir de la segunda pestaña, la información vendrá organizada en una plantilla básica dividida en dos partes irregulares entre sí. De este modo, la parte izquierda de la pantalla, significativamente más amplia, queda destinada al cuerpo de texto informativo referente al contenido de cada pestaña; en el sucinto lado derecho aparecen dos cuadros de texto fijos: el primero, informa de

las entradas más recientes del blog (*Recent Posts*) e inmediatamente debajo, otro informa de la pertenencia al proyecto europeo Horizon2020 (*Supported by*).

La segunda pestaña, *About*, es quizá la más compleja, pues cuenta con una pantalla principal a modo de resumen y un menú desplegable en la que aparecen dos páginas anidadas: *Research* y *Services*. Si pinchamos directamente en la pastilla *About* de la barra de navegación, se nos redirige a una pantalla principal en la que se nos explica, esquemáticamente, su filosofía tripartita[5]: investigación (*Research*), los servicios que ofrece (*Services*) y la red de investigación que promueve a través del uso, la investigación y el desarrollo de dichos servicios entre bibliotecas, archivos, personal investigador de Humanidades y Ciencias Sociales, así como entre el público general y los voluntarios, interesados en dichas áreas del conocimiento. Para completar la información hemos de ir a la pestaña *Network*, de la que hablaremos más adelante, aunque se hace raro este desglose a medias en que la información queda tan esparcida por la web.

Sin salir de la pestaña *About*, vamos a la subsección *Research* a través del menú desplegable de la barra de navegación. En ella se nos explica la base de esta investigación: el reconocimiento de patrones, el análisis de documentos digitalizados, el procesamiento de lenguaje natural de textos manuscritos y el entrenamiento y aprendizaje del software para dicho procesamiento de datos. El proyecto promueve la investigación a través de una red de colaboradores y encuentros en los que se reconoce la importancia de esta línea de I+D; dos pruebas de ello son la *International Conference on Frontiers in Handwriting Recognition* (ICFHR)[6], celebrada en 2016, y la *International Conference on Document Image Analysis and Recognition* (ICDAR)[7] de 2017. Por último, se incide en la importancia de la accesibilidad. En consecuencia, sus publicaciones son de acceso abierto, sus datos de investigación pueden consultarse en el repositorio Zenodo y su software está disponible en código abierto a través de GitHub.

Pasemos ahora a la segunda subsección de *About*, *Services*. En ella, se nos introduce al producto estrella de READ: Transkribus. Esta plataforma es la herramienta que hace posible los objetivos del proyecto: reconocimiento, enriquecimiento y búsqueda de material de archivo. En esta subsección se nos presenta un resumen de Transkribus: número de usuarios registrados, interfaces disponibles, servicios que ofrece el programa y la sustentabilidad. Todo ello aparece detallado de manera más pormenorizada que en la pestaña correspondiente (Transkribus), lo que no deja de ser sorprendente. Si es interesante, empero, señalar que solo desde aquí, a través de un enlace insertado en la última oración de la página, se puede acceder al rincón dedicado al futuro del proyecto: READ-COOP, una fundación aún en desarrollo, basada en la

[5] Según se anuncia en el portal del proyecto, "READ is an e-Infrastructure project funded by the European Commission and combines research, services and network building".
[6] Accesible desde: http://www.nlpr.ia.ac.cn/icfhr2016/.
[7] Accesible desde: http://u-pat.org/ICDAR2017/index.php.

cooperativa europea SCE[8], con la que se asegura seguir desarrollando y mejorando Transkribus una vez haya finalizado Horizon 2020.

Seguimos navegando por la web de READ. La tercera pestaña, *Network*, expone, de nuevo, el objetivo del proyecto: construir una plataforma que reconozca y transcriba documentos manuscritos históricos y sea capaz de buscar en ellos de manera automatizada. Uno de los pilares en los que sustenta dicha mejora de la accesibilidad del material manuscrito es la red de investigadores gracias a la cual se potencia el uso compartido y el intercambio de datos y recursos. Los catorce miembros interesados en el uso y desarrollo del proyecto y que consolidan esta red son, entre otros, la Universidad de Innsbruck, la Universitat Politècnica de Valencia, *Naver Labs* Europa (sita en Francia), el Centro Nacional de Investigación Científica Demokritos de Grecia o los Archivos nacionales de Finlandia. Además de este grupo de catorce centros, cuenta con un memorándum de entendimiento (MoU)[9] firmado por archivos y centros de investigación repartidos por todo el mundo.

La siguiente pestaña, *News*, hace las veces de blog y recopila las publicaciones que informan de los avances, las actualizaciones y los nuevos acuerdos firmados desde enero de 2016, momento en que READ tuvo su presentación en sociedad, junto con co:op[10], en la conferencia *Technology meets Scholarship, or How Handwritten Text Recognition will Revolutionize Access to Archival Collections*. Desde entonces, este blog de READ ha servido para sacar a la luz aquellos puntos en común entre archivistas, investigadores de Humanidades Digitales y especialistas en tecnologías de la información. Seguida a esta, la pestaña *Events* informa de los encuentros pasados o futuros en los que el HTR ha sido o será uno de los temas tratados en dichas jornadas.

Por fin, en sexto lugar, encontramos la pestaña de Transkribus. Parece extraño que, siendo este proyecto la cristalización de una gran parte de la investigación de READ, la pestaña para acceder a la información de la plataforma sea de las últimas en aparecer en el menú. No obstante, si bien en *Services* se nos adelantaban algunos datos, es aquí donde el equipo investigador que hay detrás de las siglas READ ha volcado toda la información aun cuando resulta algo sintética. Un breve párrafo de apenas cuatro líneas nos presenta el proyecto; el enlace a la web oficial de Transkribus[11] aparece por duplicado insertado en la primera palabra y en la última. Tras este, varias listas en las que se enumeran, todo debidamente hipervinculado, herramientas, tutoriales y numerosos manuales de instrucciones para iniciarse en el programa o transcribir y enriquecer textos. Solo al final de la página encontramos, medio escondido, el valioso enlace que nos remite a la Wiki[12] dedicada a Transkribus.

[8] Para saber más acerca de la SCE: https://bit.ly/2BaXdfS.
[9] Por sus siglas en inglés: Memorandum of Understanding.
[10] Para saber más acerca de este proyecto europeo, siga el enlace: https://coop-project.eu/.
[11] Accesible desde: https://transkribus.eu/Transkribus/.
[12] Accesible desde: https://transkribus.eu/wiki/index.php/Main_Page.

Figura 2. Logotipo de Transkribus con el wolpertinger adaptado de Albrecht Dürer de 1507.

En *Publications* se presentan cuatro enlaces tras los que podemos encontrar un sinfín de información relacionada con READ y HTR. En primer lugar, publicaciones derivadas del trabajo que saca adelante el equipo de investigación READ sobre reconocimiento de patrones, análisis de documentos digitalizados o procesamiento de lenguaje. Le siguen un enlace a *ScripNet*, la plataforma que alberga certámenes sobre HTR, análisis de diseño y tecnologías relacionadas; otro enlace a Zenodo, el repositorio en el que se encuentran, totalmente accesibles, todos los datos de investigación para las publicaciones READ, y, por último, el tercer enlace nos conduce a una lista de resultados según ocho grupos de tareas, todos debidamente enlazados a las respectivas memorias de cada periodo de actuación.

Por último, en la pestaña de C*ontact* encontramos todos los datos del coordinador del proyecto (el doctor Günter Mühlberger, de la Universidad de Innsbruck), así como de otros miembros READ ubicados en Finlandia (Maria Killo, del National Archives), Suiza (Tobias Hodel, del State Archives) o Reino Unido (Louise Seaward, del University College). También encontramos un email para consultas generales y el enlace a la Wiki de Transkribus, así como a los perfiles en redes sociales del proyecto (Twitter, Facebook y YouTube). El clásico icono de la lupa esconde el buscador de la página web y cierra el menú de navegación.

3. CONCLUSIONES

El proyecto READ persigue el reconocimiento y el enriquecimiento de documentos de archivo. A la vez que crea nuevos estándares de HTR, potencia la localización de términos clave y la identificación de patrones. Con ello, se pretende entrenar y perfeccionar el producto al que se consagra el proyecto: Transkribus. El hecho contrastado de que Transkribus permite que el material de archivo sea más accesible y la fundación READ-COOP asegura la supervivencia de

la plataforma gracias a su uso en archivos y centros de investigación una vez finalizado el periodo sufragado por la UE (2016-2019).

El gran trabajo de investigación y desarrollo del proyecto READ viene avalado por una plataforma de transcripción de manuscritos cada vez mejor entrenada gracias a la explotación de todos sus usuarios (en diciembre de 2018, más de 16.000) de forma centralizada. No obstante, se encuentran ciertas posibles mejoras en la experiencia del usuario en su web https://read.transkribus.eu/. En primer lugar, se aprecia que la información aparece, en ocasiones, dispersa o duplicada de manera injustificada. Quizá las páginas de *Network* y *Events* podrían aparecer aglutinadas en una sola pestaña del navegador con menú desplegable o, sino juntas, al menos sí una junto a la otra, como secuencia lógica en un orden de páginas con contenido similar (por un lado, las redes de profesionales que colaborar para desarrollar Transkribus o diferentes herramientas relacionadas con el HTR –*Network*–; por otro, las conferencias donde dichos profesionales se reúnen con tal fin). Algo similar ocurre con *News* y *Publications*. Bien es verdad que el primero incluye textos más breves y no siempre de carácter académico, sino informativo, mientras que el segundo se caracteriza por recoger publicaciones con rigor académico y de investigación. Quizá pueda aplicarse la misma solución que proponíamos unas líneas más arriba: ordenar las pastillas de la barra de navegación según la materia que se trata en cada página o aglutinarlas en una sola pestaña con menú desplegable. Mientras, secciones como Transkribus *Learn* o READ-COOP, en las que se tratan asuntos de verdadero peso para el equipo de investigación READ, son de difícil acceso, pues apenas cuentan con un enlace visible en toda la web. El hecho de que en la subsección *Services* se hable de manera más pormenorizada de los servicios y la proyección de Transkribus que en la página dedicada a la plataforma no deja de ser un hecho remarcable que, a nuestro entender, debería reformularse.

Para finalizar, es de extrañar que, siendo un proyecto financiado por la UE, en el que trabajan profesionales de todo el mundo, la web de READ solo esté disponible en inglés y no se tengan en cuenta otras lenguas oficiales, siquiera mayoritarias, de la UE.