Proyecto Carabela: un método revolucionario para la investigación de naufragios en archivos históricos basado en la inteligencia artificial

Enrique Vidal Ruiz
PRHLT. Universitat Politècnica de València

Carlos Alonso Villalobos
Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico
(c.alonso.sa@gmail.com)

Verónica Romero Gómez
PRHLT. Universitat Politècnica de València

Vicent Bosch Campos
PRHLT. Universitat Politècnica de València

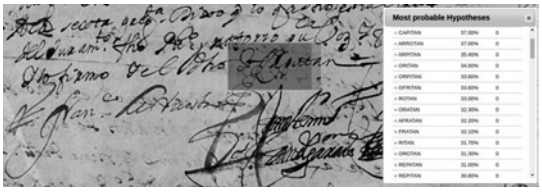
Lourdes Márquez Carmona
Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico

María del Carmen Orcero
Colaboradora del Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico

David Garrido Romero
Colaborador del Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico

Mili Jiménez Melero
Colaboradora del Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico

Data de recepció: 5 de novembre de 2020
Data d'acceptació: 9 de desembre de 2020
DOI: https://doi.org/10.51829/Drassana.28.647



■ RESUMEN

Durante los últimos veinte años, el Centro de Arqueología Subacuática (CAS-IAPH) ha desarrollado un amplio programa de documentación y recopilación de información sobre naufragios históricos en aguas andaluzas a través de los fondos de archivos públicos y privados. En la actualidad, cuenta con cerca de 2.000 referencias que van a permitir la mejora del conocimiento, protección y preservación de ese patrimonio tan amenazado por el expolio. Con el fin de mejorar la eficacia de este proceso, entre 2018 y 2019, ha desarrollado, junto con el Centro PRHLT de la Universidad Politécnica de Valencia, un proyecto innovador (el proyecto Carabela) que confirma la eficacia de las técnicas de reconocimiento de texto manuscrito (HTR) para localizar y caracterizar documentos digitales en grandes fondos de archivos sin apenas intervención humana, con el consecuente ahorro de tiempo y recursos en la investigación y/o gestión de estos fondos.

Palabras clave: arqueología subacuática, reconocimiento de textos manuscritos, aprendizaje automático, inteligencia artificial, gestión de archivos, investigación documental.

Carabela Project: a revolutionary method for shipwreck research in historical archives based on Artificial Intelligence

■ ABSTRACT

The Underwater Archeology Center (IAPH) has developed, over the last 20 years, an extensive program of documentation and compilation of information on historical shipwrecks in Andalusian waters through public and private archives. At present it has about 2000 references that will allow the improvement of knowledge, protection and preservation of that heritage, which is so much threatened by plunder. In order to improve the efficiency of this process, between 2018 and 2019, it has developed, together with the PRHLT Center of the Polytechnic University of Valencia, an innovative project (Carabela Project) that confirms the effectiveness of handwritten text recognition techniques (HTR) to locate and characterize digital documents in large archives with little human intervention, with the consequent saving of time and resources in the research and / or management of these collections.

Keywords: Underwater Archeology, Handwritten Text Recognition, Machine Learning, Artificial Intelligence, Archives management, Documentary research.

■ INTRODUCCIÓN

Hasta la invención del ferrocarril y el avión, el barco, en sus diferentes tipologías, fue el medio más utilizado para el transporte de mercancías y personas, por su rapidez y economía. Muchas de esas embarcaciones nunca llegaron a su destino por diferentes motivos: fallo del piloto, error en la estiba de la carga, ataque enemigo… y, principalmente, las condiciones meteorológicas adversas.

Cuando una embarcación naufraga se generan dos tipos de registros históricos, pero, lamentablemente, no siempre se han conservado hasta nuestros días.

- Registro material: conformado por los restos de la embarcación y/o su cargamento, permite a los arqueólogos descubrir, entre otras cosas, las posibles causas del naufragio, la cronología, el tipo de embarcación, la adscripción cultural, la nacionalidad, el cargamento, la ruta que seguía…, además de numerosos detalles sobre la vida a bordo y la naturaleza material de la estructura naval y la carga. El uso de técnicas geofísicas y de robótica subacuática ha contribuido a favorecer la localización de estos restos y su registro e inspección, y hace accesibles incluso los pecios que se encuentran a grandes profundidades y en condiciones adversas.
- Registro documental: compuesto, en ocasiones, por miles de páginas generadas desde diferentes perspectivas e intereses (información y control, rescate, reclamación a seguros, judicial, etc.), permite reconstruir, más allá de la información material, aspectos y detalles relativos a la historia del barco, el armador o la tripulación, y definir las mercancías transportadas, a sus propietarios, los seguros, los precios, etc.

Pero un naufragio es, ante todo, una catástrofe de naturaleza socioeconómica ante la que se reaccionaba en función de las posibilidades de rescate de la tripulación y la carga[1]. Supone la pérdida de vidas humanas, lo que genera un gran volumen de documentos asociados a las reclamaciones de familiares de los desaparecidos, interesados por conocer lo sucedido y por reclamar sus derechos como herederos (cartas, peticiones, informes, testamentos, etc.).

Igualmente, un naufragio es una tragedia económica que pone en marcha una compleja maquinaria de reclamaciones entre particulares, empresas y compañías aseguradoras, al objeto de determinar las causas y definir, en su caso, culpabilidades, pagos e indemnizaciones, y/o proceder a su rescate (informes, memoriales, autos, órdenes de pago, contratos con compañías de buceo, etc.).

Una riquísima fuente de información única, compuesta por millones de expedientes y documentos custodiados (según su naturaleza) en archivos, cartotecas y hemerotecas, públicos o privados. En su mayoría, estas colecciones no disponen de instrumentos de descripción adecuados (inventarios y catálogos), lo que dificulta considerablemente localizar la información de interés a sus usuarios. Solo unos pocos, los que disponen de más recursos humanos y/o económicos, están digitalizados y cuentan con herramientas que permiten realizar búsquedas, usando para ello exclusivamente los criterios cronológico, onomástico, toponímico y/o de materias normalizadas, definidos siempre a elección del archivero, lo que obliga a los investigadores a invertir muchas horas de trabajo para acceder a los expedientes o documentos que desean encontrar.

■ LA INVESTIGACIÓN SOBRE NAUFRAGIOS HISTÓRICOS: EL CASO DEL CAS-IAPH

Desde que se creara en 1998, el CAS-IAPH tiene como principal objetivo desarrollar acciones para mejorar el conocimiento, conservación y protección del patrimonio arqueológico subacuático de Andalucía mediante el desarrollo de la Carta Arqueológica Subacuática. En este sentido, el Área de Documentación Formación y Difusión del CAS-IAPH ha venido trabajando durante estos años en un amplio proyecto de documentación encaminado a localizar y sistematizar fuentes de información relativa a naufragios históricos[2]. Para ello, ha utilizado fuentes de doble naturaleza (textual y cartográfica) que, además, a través de herramientas de información geográfica, permiten conocer, interpretar y reconstruir los paisajes costeros y portuarios.

Al inicio de su labor, el CAS-IAPH tenía referencia de solo 377 naufragios por trabajos de otros autores[3]. Tras veinte años de investigación en archivos, esta cifra se ha ampliado a 1.384 (agosto de 2020). A ellos habría que sumar la referencia de otros 3.000 documentados indirectamente por fuentes primarias y secundarias en aguas internacionales o de otros países (**figura 1**).

Esta estrategia, acompañada de una sistemática labor de reconocimiento y estudio arqueológico de yacimientos por parte del Área de Intervención del CAS-IAPH, permitió elevar a la Consejería de Cultura de la Junta de Andalucía una propuesta de zonificación específica para la protección legal de numerosas zonas y/o espacios arqueológicos en aguas andaluzas[4]. Una estrategia que fue reconocida por la UNESCO en 2017 como modelo de "buenas prácticas" en desarrollo de las directrices de la Convención para la Protección del Patrimonio Arqueológico Subacuático[4].

Para la gestión de este elevado volumen de datos se generó en 2002 un modelo pionero de herramienta SIG denominado *SIGNauta* que está adaptado específicamente a las necesidades de gestión de este patrimonio[6]. Sin embargo, la crisis generalizada vivida especialmente a partir de 2010, vino a recortar drásticamente los recursos humanos disponibles para el desarrollo de esta laboriosa y lenta labor de documentación y sistematización de información, cada vez más accesible en línea. Fue en ese momento cuando, con la intención de darle continuidad, desde el CAS-IAPH se planteó la hipótesis de trabajo que, con el tiempo, daría origen al proyecto Carabela: ¿sería posible, a través de las innovadoras técnicas de la inteligencia artificial, realizar búsquedas a texto completo en un fondo documental de manuscritos de gran volumen utilizando términos de búsqueda no normalizados (sim-

Figura 1. Naufragios históricos registrados en bases de datos del CAS-IAPH (año 2020)

ples y complejos) sobre una temática de interés? Algo similar a lo que se consigue aplicando técnicas OCR a documentos impresos o mecanografiados.

■ EL PROYECTO CARABELA: ORIGEN

El origen del proyecto Carabela se encuentra precisamente en el contexto que acabamos de describir: la necesidad de agilizar la difícil y compleja tarea de identificar y localizar, entre los millones de documentos manuscritos digitalizados de archivos, aquellos que hacen referencia específicamente a naufragios históricos. Cientos de miles de embarcaciones cuyos restos debemos identificar y proteger para garantizar su preservación futura. En ocasiones, transportaban mercancías de gran valor económico (oro, plata, piedras preciosas, obras de arte, etc.) que podían ser de interés para los cazatesoros. Así sucedió, entre otros, en el mediático caso de Odyssey Marine Company y el expolio de la fragata *Nuestra Señora de las Mercedes*, naufragada en 1804 frente a las costas del sur de Portugal, identificada primeramente por el CAS-IAPH en un proyecto de colaboración con los Cuerpos de Seguridad del Estado[7].

Con este objetivo, los Centros de Investigación de Reconocimiento de Formas y Tecnología del Lenguaje Humano (PRHLT, en inglés), de la Universidad Politécnica de Valencia (UPV), y de Arqueología Subacuática, perteneciente al Instituto Andaluz del Patrimonio Histórico (CAS-IAPH), abordaron el reto de diseñar conjuntamente un proyecto que, basado en el uso de técnicas de reconocimiento de textos manuscritos (en inglés, *handwritten text recognition* [HTR]), permitiese agilizar el proceso de localización e investigación selectiva de documentación manuscrita en fondos digitalizados de archivos, de modo que ahorrara tiempo y recursos y ayudara, además, a sistematizar y categorizar dicha documentación de manera automatizada.

Ideado en 2016[8], el proyecto Carabela se desarrolla entre 2018 y 2019 gracias a la financiación de la Fundación BBVA, concedida a través de la convocatoria de Humanidades Digitales[9]. Sus resultados han sido sorprendentes y han abierto nuevas e insospechadas perspectivas a la labor de investigación documental. Unas posibilidades que no se restringen solo a la documentación de naufragios históricos, sino que pueden aplicarse a cualquier tipo de búsqueda temática y fondo documental, independientemente de su volumen, idioma y tipo de letras.

■ PROCESADO DE TEXTOS MANUSCRITOS Y SISTEMAS INTELIGENTES

El reconocimiento de formas (RF) es una disciplina científica clásica, con la que están estrechamente vinculadas otras más recientes como el aprendizaje automático (AA) y, en general, los sistemas inteligentes (SI).

Una de las aplicaciones más antiguas de RF es el reconocimiento óptico de caracteres (en inglés, *optical character recognition* [OCR]). Las tecnologías tradicionales de RF para OCR se desarrollaron en los años sesenta del pasado siglo y pronto dieron lugar a sistemas prácticos de gran utilidad para la transcripción e indexación de documentos de texto impreso o mecanografiado[10]. Las tecnologías OCR requieren una separación previa de los caracteres individuales que componen cada palabra del texto, lo que resulta relativamente sencillo para texto impreso o mecanografiado gracias al espacio en blanco que siempre separa los sucesivos caracteres. Esta misma razón hace prácticamente inviable aplicar un sistema de OCR para transcribir un texto manuscrito cursivo, en que la separación entre caracteres es casi inexistente o inconsistente, e incluso la separación entre palabras es a menudo confusa o caprichosa.

El texto manuscrito también ha sido objeto de amplios estudios en RF y AA. Pero solo en tiempos relativamente recientes han empezado a encontrarse aproximaciones metodológicas (mucho más complejas que las técnicas de OCR) que apuntan hacia soluciones prácticas para la transcripción automatizada de texto manuscrito sin restricciones. A las tecnologías que se están desarrollando actualmente en esta dirección se las suele denominar por sus siglas en inglés, HTR (*handwritten text recognition*)[11].

Del manuscrito a la imagen y a su interpretación

Hay que tener en cuenta que, para un lector humano, la lectura de texto escrito a mano generalmente requiere habilidades cognitivas complejas: para poder interpretar la información contenida en una imagen de texto manuscrito en forma de una secuencia de caracteres y de palabras, es frecuentemente necesario *entender* lo que se está leyendo. El contexto de cada carácter y palabra suele ser crucial para poder discernir cuál es cada palabra escrita que se está viendo en la imagen, y así saber cuáles son los caracteres que la forman[12].

Las tecnologías HTR se basan en métodos holísticos, que lejos de analizar cada carácter por separado (lo que por otra parte sería imposible) tratan de aproximar el complejo proceso cognitivo humano que acabamos de comentar. Es por esta razón por la que, a diferencia del OCR, un sistema HTR puede considerarse propiamente un *sistema inteligente*.

De la transcripción a la indexación probabilística

Los sistemas HTR que se están desarrollando actualmente son capaces de transcribir con precisión útil imágenes de manuscritos *sencillos* (caligrafía simple y uniforme, correcta conservación del documento y buena calidad fotográfica). Pero muchos textos históricos suelen mostrarse muy esquivos para estos sistemas. Por eso, en el proyecto Carabela se descartó desde el principio la transcripción automatizada, o incluso semiautomática (asistida)[13]. Nuestra apuesta fue por otra tecnología recientemente desarrollada por el Centro PRHLT denominada indexación probabilística (en inglés, *probabilistic indexing* [PrIx])[14].

Para cada imagen de texto, se crea una especie de *mapa de calor de palabras*. Cada píxel de este mapa indica la mayor o menor probabilidad de que ese píxel forme parte de una o, generalmente, de muchas posibles palabras o secuencias de caracteres plausibles. Para una imagen típica de las que hemos procesado en el proyecto Carabela, el índice probabilístico que representa este mapa contiene alrededor de 4.000 hipótesis de palabras o secuencias de caracteres posiblemente escritas en la imagen, con sus correspondientes probabilidades y posiciones en la imagen. Como en una imagen suele haber alrededor de 200 palabras realmente escritas, esto corresponde a una *densidad media de indexación* de unas 20 hipótesis por cada palabra real.

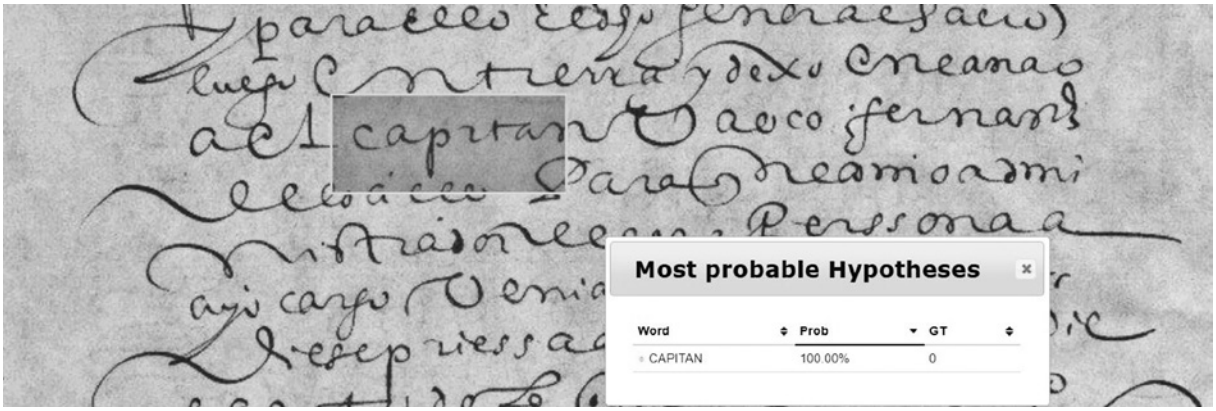
Es importante entender que un índice probabilístico de una imagen de texto es algo radicalmente diferente de una transcripción automática de esa imagen de texto. Para empezar, la transcripción sería algo mucho más simple y pequeño: alrededor de 1.000 bytes por imagen (200 palabras, a un promedio de 5 bytes por palabra). En comparación, el índice probabilístico correspondiente requiere alrededor de 60.000 bytes (4.000 hipótesis, a un promedio de 15 bytes por hipótesis). Y lo más importante: los errores de una transcripción automática son *irreversibles*.

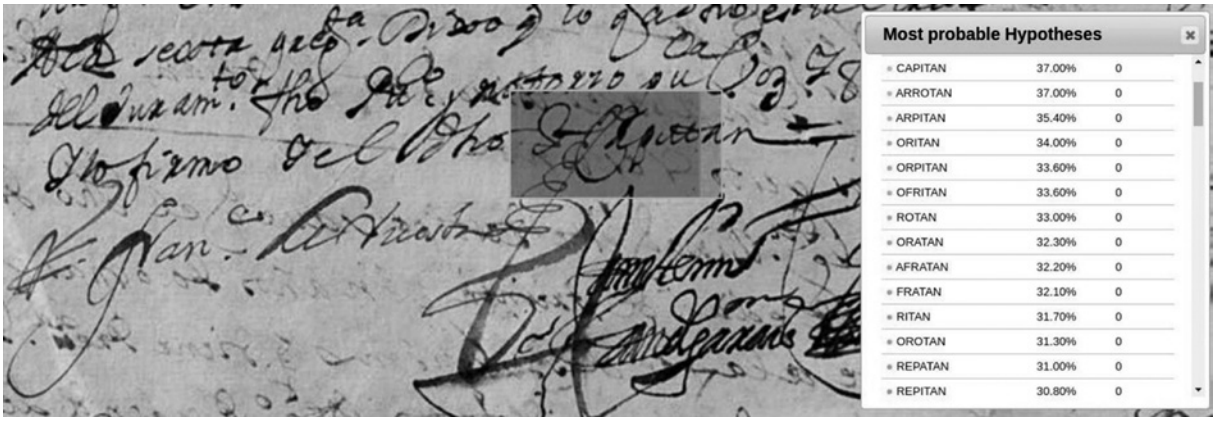
En la mayoría de documentos que se han procesado en Carabela, el 30% de las palabras (y a veces muchas más) que se obtienen mediante transcripción automática son erróneas. Si se usaran solo esas palabras para indexar la colección, los resultados de la mayoría de búsquedas de información serían decepcionantes. Por el contrario, un índice probabilístico puede contener muchas hipótesis de cada posible palabra. Si una palabra está claramente escrita, sin ambigüedades lingüísticas y la imagen es de buena calidad, el número de hipótesis que se indexan es muy bajo; quizás una sola hipótesis en muchos casos. Sin embargo, en partes deterioradas de documentos, con tipos de escritura complejos y/o ambigüedades lingüísticas (causadas, por ejemplo, por el uso de abreviaturas y/o arcaísmos), se pueden llegar a indexar hasta varios millares de hipótesis por cada posible palabra real. Las **figuras 2** y **3** ilustran este concepto. Gracias a esta *densidad adaptativa de indexación*, es posible finalmente detectar información textual precisa, incluso en las peores condiciones de la documentación analizada.

En resumen, los índices probabilísticos tratan de preservar la incertidumbre inherente a la interpretación como texto de los trazos que se observan en las imáge-

Figura 2. En una imagen de buena calidad, una palabra claramente escrita y sin ambigüedades lingüísticas se indexa con baja densidad (una sola hipótesis para la palabra *Capitán*, en este ejemplo)

Figura 3. En situaciones complejas, la misma palabra que en el ejemplo de la figura 2 queda indexada con cientos de hipótesis. En este ejemplo, se han marcado con rectángulos de colores dos de estas posibles interpretaciones de la imagen: *CAPITAN* y *REPITAN*





nes; de esta forma se evita que se pierdan posibles interpretaciones que quizás puedan parecer poco plausibles a primera vista, pero que pueden ser justamente las que interesen cuando se busque información en esas imágenes.

■ OBJETIVOS DEL PROYECTO

En el proyecto Carabela se ha trabajado en dos líneas con objetivos bien definidos:

- Valorar la eficacia de diversos métodos de reconocimiento de textos manuscritos (HTR) y de indexación probabilística (basada en técnicas de detección de palabras clave, o *keyword spotting* [KWS]), para identificar y localizar información textual en documentos de interés de colecciones masivas perteneciente a los siglos XV-XIX.
- Analizar la posibilidad de uso de las técnicas de indexación probabilística para caracterizar la documentación histórica relativa a naufragios históricos teniendo en cuenta el riesgo de un uso indebido por empresas de cazatesoros.

Con el primer objetivo se quería continuar profundizando en la línea que hasta el momento venía desarrollando el Centro PRHLT, aceptando el reto de enfrentarse a la colección de documentos más compleja hasta el momento. La complejidad radicaba no solo en la variedad de tipos de letras presentes, muy superior a la de cualquier otro experimento abordado hasta el momento con técnicas HTR, sino también en las dificultades que presentaba la colección: gran cantidad de abreviaturas, palabras arcaicas, manchas y transferencias de tintas entre las caras del documento, mal estado de conservación, etc., además de mala calidad de digitalización.

Con el segundo objetivo buscábamos ir más allá y desarrollar técnicas de análisis y clasificación de documentos (tanto individualmente como formando expedientes) que ayudasen a identificar, de modo automatizado, aquellos que específicamente interesan para nuestro objeto de estudio entre cientos de miles de documentos.

Unas técnicas que, además, permitiesen detectar y diferenciar aquellos documentos que tratan sobre embarcaciones que transportaban mercancías de valor (oro, plata, joyas, caudales, etc.) y que, por tanto, pudieran ser de interés para su expolio por cazatesoros.

Para ello se definieron las tres categorías de documentos que debíamos ser capaces de detectar:

- *Expedientes rojos*: sobre embarcaciones naufragadas que transportaban una carga de interés para los cazatesoros.
- *Expedientes amarillos*: sobre embarcaciones naufragadas con un cargamento de escaso interés para los cazatesoros.
- *Expedientes verdes*: sobre otros temas, sin hacer referencia a embarcaciones naufragadas.

■ METODOLOGÍA Y FASES DE TRABAJO

La formación de la colección (siglos XV-XIX)

Para alcanzar los objetivos planteados, debíamos partir de crear una colección experimental sobre temática comercial marítima que, en los ámbitos diplomático, paleográfico y de estado de conservación, fuera representativa de los fondos documentales propios de un archivo hispanoamericano de los siglos XV-XIX. Inicialmente, en función de las posibilidades que ofrecía la subvención recibida, se pensó en una colección de 150.000 imágenes digitales procedentes del Archivo General de Indias, en Sevilla, y el Archivo Histórico Provincial de Cádiz, ciudades ambas puerto de salida y entrada de dicho comercio. Una colección que debía albergar expedientes sobre naufragios históricos.

Conseguir la documentación no fue tarea fácil. El Archivo Histórico Provincial de Cádiz (AHPC) se mostró dispuesto desde el primer momento colaborativo y cedió con fines exclusivos de investigación 94.546 imágenes[15]. Pero acceder a las del Archivo General de Indias (AGI) fue más problemático, de modo que optamos por utilizar aquellas imágenes que se encontraban alojadas en el Portal de Archivos Españoles (PARES) del Ministerio de

Cultura y Deportes y podían descargarse en digital[16]. Si bien nuestro objetivo inicial era trabajar con imágenes de alta calidad (300 ppp), las descargadas de PARES solo alcanzaban los 125 ppp[17]. Ello nos obligó a enfrentar dos interesantes y novedosos retos no afrontados antes por otros proyectos: trabajar con imágenes de baja resolución (125 ppp) que, además, presentaban falta de nitidez y contraste de digitalización o tintas aclaradas con el paso del tiempo. Superar estos retos fue finalmente un logro interesante, dado que estas condiciones extremas son las que se encuentran frecuentemente en colecciones custodiadas en archivos históricos.

A través de la plataforma PARES se seleccionaron expedientes del AGI con diferentes criterios sobre los contenidos, según la información disponible en los correspondientes metadatos del expediente (naufragio, pérdida, buceo, buzo, flota, azogue, etc.). Así se descargaron unas 45.000 imágenes. En el proceso de revisión, organización y creación de nombres de las imágenes del AHPC y del AGI se comprobó que algunas estaban duplicadas, de modo que finalmente la colección quedó compuesta por 125.311 imágenes, divididas en dos subcolecciones (**tabla 1**):

- Subcolección AGI: con 328 expedientes y un total de 30.765 imágenes de documentos generados en el contexto de la actividad de la Casa de la Contratación.
- Subcolección AHPC: con 50 protocolos notariales y un total de 94.546 imágenes, entre ellas testamentos, seguros marítimos, cartas de pago, etc., formalizados ante notario por los herederos de marineros fallecidos que reclamaban sus bienes, o por comerciantes que demandaban a las aseguradoras las indemnizaciones a que tenían derecho por la pérdida de las mercancías de su propiedad durante un naufragio.

Tratamiento de la documentación digital

Como se ha comentado anteriormente, la indexación probabilística puede hacerse sin una detección previa de las líneas o zonas de la imagen donde se encuentran las palabras[18]. No obstante, el proceso de indexación puede acelerarse considerablemente mediante un proceso previo en el que se detectan, al menos toscamente, las posibles líneas de texto existentes en la imagen.

En Carabela, este preproceso simplificado se llevó a cabo mediante un sistema de análisis de maquetación similar al propuesto por Quirós, Toselli y Vidal[19], entrenado mediante unas pocas imágenes de la colección con las líneas de texto marcadas manualmente.

Entrenamiento del sistema de indexación probabilístico

Esta es, sin duda, la fase más delicada de todo el proceso: el entrenamiento del sistema, para que sea capaz de reconocer los diferentes tipos y variantes de letras que contiene la documentación. No solo es importante la cronología (del siglo XV al XIX: gótica, cortesana, procesal, etc.), sino también la variedad de escribanos ("manos"), las condiciones de escritura y el tipo de material empleado: herramienta, soporte, tinta, etc. (**figura 4**).
Particularmente importantes son las abreviaturas y los arcaísmos, que en la colección considerada se usan con enorme frecuencia, y además no siguen reglas claras y homogéneas. Para ello se generó una pequeña colección de entrenamiento (*ground-truth*) compuesta por 557 imágenes de documentos cuidadosamente seleccionados para que fuesen representativas de la colección (tipos de letras, transferencias de tintas entre las caras, manchas, roturas, falta de soporte, etc.).

Este grupo de entrenamiento se subdividió en cinco lotes que se fueron trabajando sucesivamente. Los dos primeros contenían una selección de imágenes extraídas del *Libro de Apeo de Purchil y Purchilejo* de 1571, cuyas imágenes fueron cedidas por el Archivo Histórico Provincial de Granada[20].

Para el tratamiento del primer lote se adaptaron los sistemas de entrenamiento de modelos ópticos y de lenguaje, y las herramientas *software* de que disponía el laboratorio PRHLT. Con ello se obtuvieron grafos de caracteres y de palabras cuyas transcripciones se modernizaron re-

Figura 4. En las primeras cuatro líneas se pueden observar variantes de la palabra *capitan*; en la inferior, formas arcaicas de los nombres *Francisco José*, *San Cristóbal* y *Jesucristo*. Todas estas palabras (junto a otros miles más) se encuentran con facilidad desde la interfaz de búsqueda de Carabela, simplemente usando las versiones modernas y no abreviadas de las palabras (sin acento): *Capitan*, *Jose*, *Cristobal* y *Jesucristo*

Tabla 1. Composición de la colección y el índice probabilístico de Carabela, con datos estadísticos extraídos de los índices. Por *spot* se entiende una palabra (o pseudopalabra), junto con el tamaño y posición de su caja de inclusión en la imagen

	AGI	AHPC	CARABELA
Número de subcolecciones (archivos)	1	1	2
Número de expediciones	328	50	378
Número de imágenes indexadas	30.765	94.546	125.311
Número total de *spots*	82.787.523	402.905.694	485.693.217
Número medio de *spots* por imagen	2.691	4.261	3.875
Número estimado de palabras	5.879.328	19.704.338	25.583.666
Tamaño estimado del léxico	331.836	233.749	566.757
Número medio estimado de palabras por imagen	191	208	204
Densidad (n.º de *spots* / n.º estimado de palabras)	**14,1**	**20,4**	**19,0**

curriendo al uso automatizado de diccionarios, si bien, en determinados casos, fueron corregidas manualmente, desarrollando las palabras representadas de forma abreviada. Los modelos de reconocimiento de lenguaje y ópticos obtenidos con el primer lote sirvieron de semilla inicial para la transcripción asistida del resto de los lotes de imágenes de entrenamiento.

Mediante observación visual se verificó que muchos de los errores se producían en imágenes con un tipo de caligrafía escasamente entrenada, por lo que, para mejorar la eficacia del sistema y alcanzar el nivel de exigencia perseguido, se fue incluyendo ese tipo de imágenes en los últimos bloques de entrenamiento.

Para facilitar el trabajo de transcripción se empleó la herramienta CATTI (en inglés, *Computer Assisted Transcription of Text Images*), desarrollada por el centro PRHLT[21]. Partiendo de una propuesta de transcripción ofrecida por CATTI para cada línea del documento, se procedía a corregir palabra a apalabra los posibles errores que se detectaban, con lo que no solo se consiguió mejorar su eficacia por el entrenamiento, sino también ahorrar tiempo y facilitar la labor de los paleógrafos.

Para conocer la evolución de la precisión conseguida por el sistema tras los sucesivos lotes de entrenamiento, véase la bibliografía de referencia en la nota 14. Aunque tras los dos primeros lotes ya se obtenían resultados útiles, se consideró conveniente completar el entrenamiento con los cuatro primeros lotes (el lote quinto se reservó para evaluación). De esta forma, finalmente se consiguió una mejora significativa con respecto a la precisión inicial.

Concluida la fase de entrenamiento[22], se reinició el sistema y se procesaron todas las imágenes del fondo a través de los nuevos modelos de entrenamiento ópticos y de lenguaje. Con ello se generaron los índices probabilísticos de la colección completa que, como veremos, constituyen el soporte esencial para dar respuesta a las búsquedas formuladas por los usuarios (**figura 5**).

Durante esta misma fase de entrenamiento se diseñaron y generaron los descriptores *semánticos* (a los que denominamos *macros*), necesarios para determinadas búsquedas complejas de uso frecuente. Para entender esta funcionalidad, imaginemos que un usuario quiere localizar, a través del sistema, documentos sobre el naufragio de una embarcación denominada *Victoria*, pero cuya tipología desconoce. Sin el soporte de los macros, se debería introducir en su búsqueda (con la ayuda de funciones booleanas) el nombre de la embarcación (*Victoria*), seguido de todos los posibles tipos de barcos de la época (nao, navío, carabela, urca, etc.), para diferenciarlo del nombre de mujer, y seguido de todos los posibles términos utilizados en la época para referirse al hecho de naufragar (hundido, perdido, a pique, etc.). En total varias decenas de términos diferentes. Con la ayuda de los macros, solo se necesitaría introducir, además del nombre, dos términos: el macro de barcos (@*barcos*, que agrupa todos los tipos de embarcaciones) y el de naufragio (@*naufragios*, que reúne todos los términos de la época sinónimos de naufragio). Igualmente se procedió con los macros @*buceo* y @*caudales*; este último agrupa términos relativos a cualquier tipo de carga valiosa.

De esta forma, utilizando los macros[23] y cualquier otro tipo de término simple (por ejemplo: el nombre de un barco, de un capitán, de un lugar, etc.) se pueden realizar gran variedad de búsquedas con ahorro de tiempo para el usuario.

Clasificación y reconocimiento

Una vez sistematizada toda la información, y tras superar la fase de aprendizaje del sistema, se creó la interfaz de búsqueda desde la que interactuar con el fondo como usuario. La base de esta son los índices probabilísticos de las imágenes de la colección. La parte más importante de la interfaz es un avanzado motor de búsqueda que da respuesta prácticamente instantánea a cualquier consulta compleja que se formule sobre cualquier colección compuesta por cientos de miles, o incluso millones, de imágenes de texto manuscrito. Entre las opciones de búsqueda avanzada soportadas se incluyen las búsquedas booleanas (AND/OR/NOT), búsquedas AND con proximidad geométrica de los términos, secuencias de palabras, búsquedas aproximadas, búsquedas usando comodines,

Figura 5. Esquema de funcionamiento
y uso del sistema

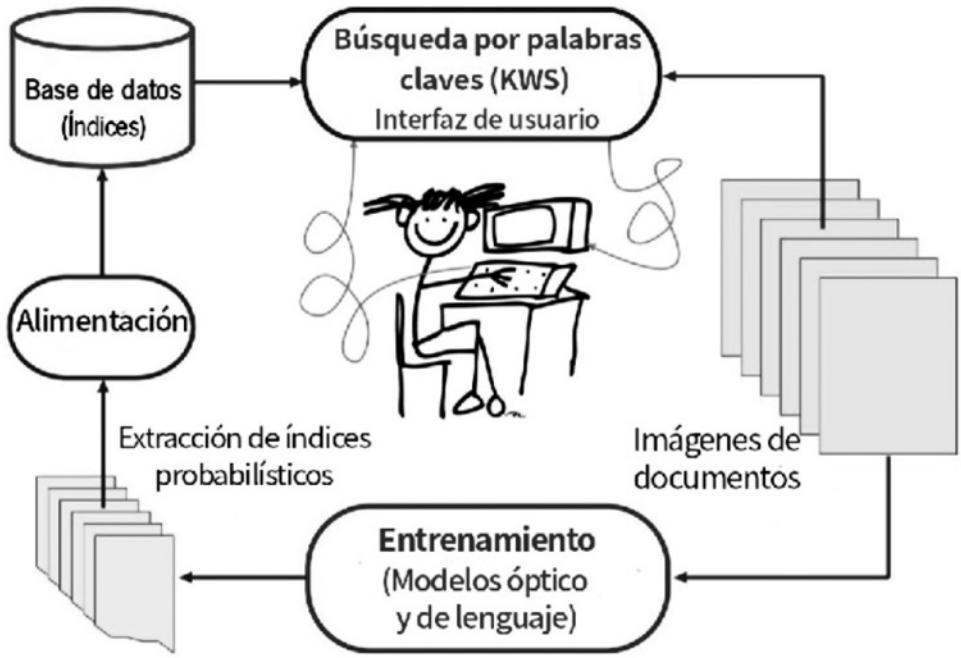


Figura 6. Interfaz de búsqueda de Carabela

etc. Esta interfaz de búsqueda esta públicamente disponible (véase la nota 9), aunque con algunas restricciones en la parte de AGI (figura 6).

Paralelamente, se perfeccionaron las técnicas de análisis y clasificación de imágenes (tanto individualmente como agrupadas en expedientes) para conseguir que el sistema fuese capaz de reconocer y clasificar todo el fondo documental conforme a las tres categorías ya definidas (expedientes rojos, amarillos y verdes). En trabajos recientes, se presentan las tecnologías de clasificación desarrolladas para abordar este complejo problema[24].

■ EVALUACIÓN

En esta sección se explican las pruebas que se llevaron a cabo en Carabela orientadas a evaluar la usabilidad de los sistemas de búsqueda y de clasificación. En los dos últimos trabajos citados en notas se reportan resultados de evaluación objetiva basada en métricas canónicas. De las dos subcolecciones (AHPC y AGI), la de mayor interés para evaluar el sistema era la del AGI. De los 328 expedientes del AGI (30.765 páginas) conocíamos que muchos contenían información sobre naufragios, pues se seleccionaron mediante búsquedas selectivas. Sin embargo, el fondo del AHPC (50 protocolos notariales, con 94.546 páginas) se seleccionó desconociendo su contenido. Una información que se ocultó a una parte del equipo hasta finalizar los trabajos de clasificación automática para evitar interferencias en el proceso.

Para evaluar la eficacia de Carabela se decidió trabajar en tres líneas:

- Eficacia para localizar e identificar palabras o términos simples de búsqueda.
- Capacidad para localizar expedientes con referencias a naufragios mediante búsquedas complejas.
- Efectividad para detectar y clasificar los expedientes denominados *rojos*, *amarillos* y *verdes*.

Era importante que el equipo en general, y en especial el encargado de abordar la evaluación automática, desconociera *a priori* las características de la documentación, por lo que se dejó para el final el análisis sistemático de su contenido. Ante la imposibilidad de abordar la lectura y clasificación de la totalidad de los expedientes, se decidió crear una muestra de evaluación abarcable y representativa, finalmente compuesta por 201 de los 328 expedientes. Cada uno de ellos se analizó, anotó y clasificó manualmente conforme a las tres categorías prefijadas (expedientes rojos, amarillos o verdes), de modo que conformaron un grupo de referencia para evaluar en todas sus variantes la eficacia del sistema.

Localización e identificación de palabras o términos simples de búsqueda

La muestra de datos para esta línea de evaluación se extrajo al azar de 40 expedientes del AGI. De cada uno se seleccionaron 4 términos entre los utilizados por los archiveros al catalogar los expedientes: 160 términos en total. De ellos, 17 no aparecían en los documentos originales, sino que los introdujo subjetivamente el archivero como descriptor temático durante el proceso de catalogación. Así, por ejemplo, encontramos el término *naufragio*, cuando en realidad el documento decía que "la nao se perdió".

Una vez eliminados, la muestra de evaluación se redujo a 143 términos, con los que se hicieron búsquedas en Carabela cambiando el nivel de confianza entre el 50% y el 0% sucesivamente. Solo 14 términos no fueron detectados; por lo general, ello fue debido a una mala digitalización, la complejidad de la grafía o el estado de conservación del documento. En resumen, el sistema fue capaz de detectar con éxito el 90,2% de los términos (tabla 2).

Tabla 2. Resultados de búsquedas por palabras o términos simples

Nivel de confianza	N.º de términos detectados	N.º de términos no detectados
50%	107	36
40%	9	27
20%	13	14

Localización de expedientes con referencias a naufragios mediante búsquedas complejas

Para facilitar la búsqueda de expedientes relacionados con naufragio de embarcaciones se diseñaron los macros @*naufragios* y @*barcos*. En estos casos, la efectividad del sistema resultó muy elevada, ya que detectó 170 de los 173 expedientes existentes en el grupo de evaluación, con un 98,2% de efectividad.

Detección y clasificación de expedientes por tipos

En esta línea, el reto era que el sistema fuese capaz de detectar y clasificar la documentación en tres tipos de expedientes: rojo, amarillo y verde. Para ello fue necesario generar, además de los macros @*naufragios* y @*barcos*, los de @*buceo* y @*caudales*.

Tras conocer los datos de búsqueda y clasificación automatizada, se procedió a analizar y clasificar manualmente los 201 expedientes de muestra, y con ello se generó la información de base que serviría para evaluar la eficacia del sistema desarrollado.

Contrastando unos y otros se pudo evaluar la eficacia de Carabela en este campo. El resultado fue sorprendente:

Tabla 3. Evaluación de la eficacia del sistema

Categoría	N.º expedientes reales	N.º expedientes detectados por el sistema	Eficacia (%)
Expedientes rojos	87	89	97,7
Expedientes amarillos	83	84	98,7
Expedientes verdes	31	28	90,3

la aplicación del proyecto había sido capaz de clasificar de manera automática el 97,7% de los expedientes rojos, el 98,7 de los amarillos y el 90,3% de los verdes (**tabla 3**).

■ CONCLUSIONES

Con el desarrollo del proyecto Carabela hemos podido confirmar que las técnicas de inteligencia artificial para reconocimiento de texto manuscrito (HTR), incluyendo nuevos conceptos como la indexación probabilística, la búsqueda semántica de información y la clasificación de imágenes no transcritas por sus contenidos textuales, son de gran utilidad para la identificación y reconocimiento de fondos documentales manuscritos digitales. Los resultados alcanzados son excelentes, tanto en detección de palabras como de expedientes, y, además, permiten clasificarlos en función del contenido textual. Sin duda, un campo innovador que abre grandes perspectivas para la investigación y gestión de grandes fondos documentales manuscritos digitalizados.

Permiten al usuario localizar entre millones de documentos aquellos de su interés (gestor, investigador o archivero) sin apenas necesidad de intervención humana, no solo sobre naufragios históricos (objeto de nuestro proyecto), sino sobre cualquier otra temática que pueda aparecer recogida en la documentación, utilizando términos modernizados y sin barrera idiomática.

Además, permite la clasificación automatizada de la documentación sin necesidad de intervención humana, lo que abre una línea de trabajo revolucionaria (por innovadora y por su inmenso potencial) para las labores de gestión y ordenación de estos fondos documentales.

Su uso agiliza enormemente la labor de investigación en archivos. Ahorra meses y años de trabajo a los investigadores, y ayuda, además, a comprender el significado o la transcripción de las palabras. En nuestro caso, basta comparar el esfuerzo realizado por el CAS durante veinte años para recuperar información de 2.000 referencias de naufragios, con el tiempo inferior a 1 minuto invertido en localizar en Carabela los 170 expedientes de su subcolección con un 98,2% de efectividad.

No hablamos de un sistema o *software* de solución única, sino de unas técnicas que deben implantarse individualizadamente y ser adaptadas a la realidad de cada colección. Un innovador desarrollo que permitiría a las instituciones que tutelan y gestionan los fondos documentales ofrecer un innovador servicio para agilizar y facilitar a los usuarios el acceso a la información con un enorme ahorro de tiempo, esfuerzo y recursos económicos.

Las técnicas de reconocimiento de textos manuscritos (HTR) e indexación probabilística basada en estrategias de detección de palabras clave (KWS) y aprendizaje profundo de redes neuronales han venido desarrollándose en el Centro PRHLT durante décadas. El proyecto Carabela ha sido una piedra angular en este desarrollo. Si bien nuestra colección facticia es de tamaño medio, las complejas pruebas realizadas con éxito demuestran que estas tecnologías se encuentran suficientemente maduras en la actualidad como para poder afrontar el mismo reto en fondos complejos de millones de documentos.

■ BIBLIOGRAFÍA

ALONSO VILLALOBOS, *et al.*, "SIGNauta: un sistema para la información y gestión del patrimonio arqueológico subacuático de Andalucía", en *PH: Boletín del Instituto Andaluz del Patrimonio Histórico*, n.º 63 (agosto 2007): 26-41.

ALONSO VILLALOBOS, Carlos, y MÁRQUEZ CARMONA, Lourdes, "Fuentes de información del patrimonio arqueológico subacuático de Andalucía. Una década de investigación documental, Arqueología subacuática española", en *Actas del I Congreso de Arqueología Náutica y Subacuática Española*, *Cartagena, 14, 15 y 16 de marzo de 2013*, Cádiz: Universidad de Cádiz, 2014, vol. 2, 91-100.

ALONSO VILLALOBOS, Carlos, *et al.*, "El uso de nuevas tecnologías para el acceso a la información histórica manuscrita en soporte digital. El Proyecto Galeón", en *Actas del V Congreso Internacional de Arqueología Subacuática IKUWA V*, Cartagena: Ministerio de Cultura y Deporte, 2016, 247-258.

BARRERE, Killian, TOSELLI, Alejandro H., y VIDAL, Enrique, "Line Segmentation Free Probabilistic Keyword Spotting andIndexing", en *Iberian Conference on Pattern Recognition and Image Analysis*, Madrid: Spinger, 2019, 201-213.

CAMPO HERNÁN, María del Pilar del, "Los archivos y la protección del patrimonio: La última comisión de la fragata de guerra *Nuestra Señora de las Mercedes*", en *La Moneda: Investigación numismática y fuentes archivísticas*, Madrid: Asociación de Amigos del Archivo Histórico Nacional y Dpto. de Ciencias y Técnicas Historiográficas y de Arqueología, UCM, 2012, 263-292.

CHAUNU, Pierre, *Sevilla y América: siglos XVI y XVII*, Sevilla: Secretariado de Publicaciones de la Universidad de Sevilla, 1983.

GARCÍA-BAQUERO GONZÁLEZ, Antonio, *Cádiz y el Atlántico (1717-1778): el comercio colonial español bajo el monopolio gaditano*, Cádiz: Diputación de Cádiz, 1988.

GARCÍA RIVERA, Carmen, y ALZAGA GARCÍA, Milagros, "La carta arqueológica subacuática de Andalucía como instrumento para la tutela de un patrimonio emergente", en *Mainake*, n.º 30 (2008): 129-143.

GARCÍA RIVERA, Carmen, y ALZAGA GARCÍA, Milagros, "La tutela del patrimonio arqueológico subacuático en Andalucía", en *Actas del V Congreso Internacional de Arqueología Subacuática IKUWA V*, Cartagena: Ministerio de Cultura y Deporte, 2016, 89-98.

HERNÁNDEZ TORNERO, Celio, *et al.*, "Indexación y reconocimiento automático de texto manuscrito", en *Cuadernos AISPI*, n.º 11 (2018): 131-144.

Libro de Apeo de Purchil y Purchilejo ([1517], s. l., 2017).

PÉREZ-MALLAINA BUENO, Pablo Emilio, *Naufragios en la Carrera de Indias durante los siglos XVI y XVII. El hombre frente al mar*, Sevilla: Universidad de Sevilla, 2015.

PINO RUIZ, Arturo del, y RODRÍGUEZ GONZÁLEZ, María Rosario, "Zonas y Servidumbres Arqueológicas: La novedosa protección del patrimonio arqueológico subacuático en Andalucía", en *PH. Boletín del Instituto Andaluz del Patrimonio Histórico*, n.º 67 (agosto 2008): 88-99.

PRIETO, José Ramón, *et al.*, "Textual-Content-Based Classification of Bundles of Untranscribed Manuscript Images", en *26th. Int. Conf. on Pattern Recognition, proceedings* (Milán, 2021), 3162-3169.

PUIGCERVER, Joan, *A probabilistic formulation of keyword spotting*, tesis doctoral (Valencia, 2018).

QUIRÓS, Lorenzo, TOSELLI, Alejandro H., y VIDAL, Enrique, "Multi-task layout analysis of handwritten musical

scores", en *Iberian Conference on Pattern Recognition and Image Analysis*, Madrid: Springer, 2021, 123-134.

ROMERO GÓMEZ, Verónica, TOSELLI ROSSI, Alejandro Héctor, y VIDAL RUIZ, Enrique, *Multimodal Interactive Handwritten Text Recognition*, Singapur: World Scientific Publishing, 2012.

SERRANO MANGAS, Fernando, *Naufragios y rescates en el tráfico indiano durante el siglo XVII*, Madrid: Siruela, 1991.

TOSELLI, Alejandro H., *et al.*, "Probabilistic multi-word spotting in handwritten text images", en *Pattern Analysis and Applications*, vol. 22, n.º 1 (2019): 23-32.

VIDAL, Enrique, *et al.*, "The Carabela Project and Manuscript Collection: Large-Scale Probabilistic Indexing and Content-based Classification", en *17th Proc. of the International Conference on Frontiers in Handwriting Recognition (ICFHR)* (Dortmund, 2020), 85-90.

■ NOTAS

1. Pablo Emilio PÉREZ-MALLAINA BUENO, *Naufragios en la Carrera de Indias durante los siglos XVI y XVII. El hombre frente al mar* (Sevilla, 2015), 270.

2. Carlos ALONSO VILLALOBOS y Lourdes MÁRQUEZ CARMONA, "Fuentes de información del patrimonio arqueológico subacuático de Andalucía. Una década de investigación documental, Arqueología subacuática española", en *Actas del I Congreso de Arqueología Náutica y Subacuática Española, Cartagena, 14, 15 y 16 de marzo de 2013* (Cartagena, 2014), vol. 2, 91-100.

3. Pierre CHAUNU, *Sevilla y América: siglos XVI y XVII* (Sevilla, 1983); Fernando SERRANO MANGAS, *Naufragios y rescates en el tráfico indiano durante el siglo XVII* (Madrid, 1991); Antonio GARCÍA-BAQUERO GONZÁLEZ, *Cádiz y el Atlántico (1717-1778): el comercio colonial español bajo el monopolio gaditano* (Cádiz, 1988), entre otros.

4. Arturo del PINO RUIZ y María Rosario RODRÍGUEZ GONZÁLEZ, "Zonas y Servidumbres Arqueológicas: La novedosa protección del patrimonio arqueológico subacuático en Andalucía", en *PH. Boletín del Instituto Andaluz del Patrimonio Histórico*, n.º 67 (agosto, 2008): 88-99.

5. Carmen GARCÍA RIVERA y Milagros ALZAGA GARCÍA, "La carta arqueológica subacuática de Andalucía como instrumento para la tutela de un patrimonio emergente", en *Mainake*, n.º 30 (2008): 129-143; Carmen GARCÍA RIVERA y Milagros ALZAGA GARCÍA, "La tutela del patrimonio arqueológico subacuático en Andalucía", en *Actas del V Congreso Internacional de Arqueología Subacuática IKUWA V* (Cartagena, 2016), 89-98.

6. Carlos ALONSO VILLALOBOS, *et al.*, "SIGNauta: un sistema para la información y gestión del patrimonio arqueológico subacuático de Andalucía", en *PH. Boletín del Instituto Andaluz del Patrimonio Histórico*, n.º 63 (agosto 2007): 26-41.

7. María del Pilar del CAMPO HERNÁN, "Los archivos y la protección del patrimonio: La última comisión de la fragata de guerra *Nuestra Señora de las Mercedes*", en *La Moneda: Investigación numismática y fuentes archivísticas* (Madrid, 2012), 263-292.

8. Carlos ALONSO VILLALOBOS, *et al.*, "El uso de nuevas tecnologías para el acceso a la información histórica manuscrita en soporte digital. El Proyecto Galeón", en *Actas del V Congreso Internacional de Arqueología Subacuática IKUWA V* (Cartagena, 2016), 247-258.

9. **http://carabela.prhlt.upv.es/es**

10. Celio HERNÁNDEZ TORNERO, *et al.*, "Indexación y reconocimiento automático de texto manuscrito", *Cuadernos AISPI*, n.º 11 (2018), 131-144.

11. Verónica ROMERO, Alejandro Héctor TOSELLI y Enrique VIDAL, *Multimodal Interactive Handwritten Text Recognition* (Singapur, 2012).

12. ROMERO, TOSELLI y VIDAL, *Multimodal Interactive Handwritten Text Recognition*.

13. ROMERO, TOSELLI y VIDAL, *Multimodal Interactive Handwritten Text Recognition*.

14. Véase al respecto Enrique VIDAL RUIZ, *et al.*, "The Carabela Project and Manuscript Collection: Large-Scale Probabilistic Indexing and Content-based Classification", en *17th Proc. of the International Conference on Frontiers in Handwriting Recognition (ICFHR)* (Dortmund, 2020), 85-90; Joan PUIGCERVER, *A probabilistic formulation of keyword spotting*, tesis doctoral (Valencia, 2018).

15. Quisiéramos resaltar el interés y apoyo que siempre mostró para con el proyecto su anterior director Manuel Cañas, lamentablemente fallecido. Un magnífico gestor y mejor persona. *Sit tibi terra levis* (Que la tierra te sea leve)..

16. **http://pares.culturaydeporte.gob.es/inicio.html**

17. Tras utilizar unas pocas decenas de imágenes, se estimó que la precisión alcanzada al usar imágenes de 125 ppp era menos de un 5% menor que la precisión alcanzable con las mismas imágenes con resolución de 300 ppp.

18. Killian BARRERE, Alejandro H. TOSELLI y Enrique VIDAL, "Line Segmentation Free Probabilistic Keyword Spotting andIndexing", en *Iberian Conference on Pattern Recognition and Image Analysis* (Madrid, 2019), 201-213.

19. Lorenzo QUIRÓS, Alejandro H. TOSELLI y Enrique VIDAL, "Multi-task layout analysis of handwritten musical scores", en *Iberian Conference on Pattern Recognition and Image Analysis. Springer* (Madrid, 2019), 123-134.

20. *Libro de Apeo de Purchil y Purchilejo* ([1517], s. l., 2017).

21. ROMERO, TOSELLI y VIDAL, *Multimodal Interactive Handwritten Text Recognition*.

22. La duración de este proceso y el número de páginas a entrenar dependen de la complejidad de la colección (variedad y dificultad de tipos de letras, abundancia de abreviaturas, problemas de conservación, etc.). En el caso de la colección de Carabela, el proceso de entrenamiento duró unos seis meses de trabajo continuado.

23. Los macros pueden ser fácilmente definidos e incorporados al sistema, aunque en el demostrador actual esta funcionalidad está disponible solo para el administrador del sistema.

24. Véanse, por ejemplo, VIDAL *et al.*, "The Carabela Project", 85-90; y José Ramón PRIETO, *et al.*, "Textual-Content-Based Classification of Bundles of Untranscribed Manuscript Images", en *26th. Int. Conf. on Pattern Recognition, proceedings* (Milán, 2021), 3162-3169.