El uso de nuevas tecnologías para el acceso a la información histórica manuscrita en soporte digital. El Proyecto Galeón

The use of new technologies to access to handwritten historical information in digital form. Galeón Project

Carlos Alonso Villalobos
Instituto Andaluz del Patrimonio Histórico. Centro de Arqueología Subacuática. Jefe del Área de documentación y transferencia
carlos.alonso.v@juntadeandalucia.es

Dr. Moisés Pastor Gadea
Pattern Recognition and Human Language Technologies research centre
Universidad Politècnica de València
mpastorg@prhlt.upv.es

Dr. Enrique Vidal
Pattern Recognition and Human Language Technologies research centre
Universidad Politècnica de València
evidal@prhlt.upv.es

Lourdes Márquez Carmona
Instituto Andaluz del Patrimonio Histórico. Centro de Arqueología Subacuática. Técnico del Área de documentación y transferencia
lourdes.marquez.carmona@juntadeandalucia.es

Resumen: la investigación histórica en archivos obliga a realizar un amplio trabajo de revisión de miles de documentos que, en muchos casos, no tienen relación con el tema de estudio, generando un importante gasto en tiempo y recursos. Para dar respuesta a este problema en relación al estudio del patrimonio arqueológico subacuático, desde el CAS-IAPH se ha ideado el Proyecto Galeón, cuyo objetivo es desarrollar soluciones innovadoras para consultar grandes conjuntos digitalizados de documentos históricos manuscritos. Actualmente no es posible la transcripción automatizada de un gran volumen de imágenes de documentos manuscritos, pero

el desarrollo tecnológico en el campo del reconocimiento formal de palabras, puede simplificar este proceso. Para ello se ha ideado un modelo teórico de Búsqueda de Palabras Claves (BPC) basado en Grafos de Palabras (GP), que, además de para el patrimonio cultural marítimo, podría utilizarse para otros temas de investigación.

Palabras clave: técnicas de reconocimiento de textos manuscritos, grafos de palabras, búsqueda de imágenes e indexación, patrimonio cultural, arqueología subacuática.

Abstract: Historical research in archives forces to realize an extensive work of reviewing thousands of documents that, in many cases, have no connection with the subject matter, generating a significant expenditure of time and resources. To address this problem in relation to the study of underwater archaeological heritage, from the CAS-IAPH has been devised the Galleon Project, which aims to develop innovative solutions to query large sets of historical documents digitized manuscripts. Nowadays It is not possible the automated transcription of a large volume of images from handwritten documents, but the development in the field of formal recognition of words, can simplify this process. For this we have developed a theoretical model to identify Keywords based on Graphs of Words (GP), which, as well as in the maritime cultural heritage, could be used for any research topic.

Key words: Handwritten text recognition technology, word-graph based keyword spotting, image spotting and indexing, cultural heritage, underwater archaeology.

Introducción

Desde la más remota Antigüedad hasta la invención del tren y el avión, el barco fue el principal medio de transporte. Por diferentes causas (accidentes durante la navegación, tormentas, huracanes, conflictos armados, etc.) muchos de ellos no llegaron a puerto, preservándose en el fondo de los océanos, ríos y lagos un gran número de pecios históricos que esconden buena parte de la historia marítima y comercial de diferentes países. En los últimos años se ha producido un gran desarrollo técnico en el campo de la exploración marina, lo que está permitiendo conocer mejor este importante legado cultural, cuyo atractivo histórico, significado, belleza y autenticidad, le convierten en un recurso más a tener en cuenta en los programas regionales de desarrollo económico.

El patrimonio arqueológico subacuático, sin embargo, conforma una de las realidades patrimoniales peor conocidas. Frente a la arqueología terrestre, ampliamente representada por multitud de proyectos y líneas de investigación, la subacuática se encuentra escasamente apoyada y desarrollada, fruto, entre otros factores, de la juventud de esta disciplina. Durante los últimos años el número de yacimientos arqueológicos subacuáticos conocidos, inventariados y protegidos ha ido en aumento, si bien este es aún escaso si atendemos al alto potencial de naufragios históricos que pone de manifiesto la investigación de las fuentes de información custodiada en archivos y bibliotecas. Unos pecios que, al igual que los inventariados y estudiados con metodología arqueológica, debemos conocer y documentar para protegerlos y, paulatinamente localizarlos, estudiarlos y conservarlos como parte de nuestro legado histórico-cultural.

Todo proyecto de investigación histórica debe fundamentarse sobre el estudio de las fuentes de información que, en función de la naturaleza de los bienes que hay que analizar y del periodo histórico al que pertenecen, pueden ser de diferente naturaleza. Recuperar la historia económica de un antiguo asentamiento del Valle del Nilo, de Mesopotamia, el Egeo, etc., pasa imprescindiblemente por el análisis multidisciplinar de las evidencias materiales

conservadas en los contextos arqueológicos. Pero también, allá donde se han conservado textos escritos en cualquier tipo de soporte, es necesaria la lectura y estudio de estas fuentes históricas, al objeto de conocer el testimonio de las personas que directamente se vinculaban con esa actividad.

Dependiendo del pasado cultural de la zona del planeta en que nos encontremos, estas fuentes de información escrita pueden remontarse a momentos más o menos antiguos. Para el caso de España, a pesar de contar con una importante tradición literaria greco-latina y una abundante información epigráfica, la información histórica que proporcionan es bastante escasa y sesgada para analizar aspectos tales como el que nos ocupa: la actividad comercial marítima en general y el estudio de los naufragios en particular.

Conocer la navegación y el comercio marítimo de cualquier zona del territorio español con anterioridad a la Baja Edad Media pasa, casi con exclusividad, por un análisis arqueológico de los restos conservados en yacimientos terrestres y subacuáticos, única fuente de información con la que cuentan los historiadores ante la escasez y parquedad de datos recogidos en las fuentes escritas, hasta ese momento más preocupadas por asuntos tales como los político-administrativos y los religiosos.

A partir del siglo XIII se produce un cambio de la situación. El desarrollo de la vida urbana y de las transacciones comerciales a gran escala fueron imponiendo cada vez más la necesidad de dejar constancia escrita de, entre otros temas, los acuerdos y negocios entre los Estados, los particulares o entre ambos, favorecida a partir del siglo XV con el desarrollo de la imprenta.

En España este punto de inflexión lo conforma la promulgación en 1503 por parte de la reina Isabel la Católica de la pragmática de Alcalá de Henares, en la que se regulaba, entre otros asuntos, el registro y conservación de toda la documentación administrativa. Desde ese momento este tipo de registros debían custodiarse encuadernados, siendo considerado este hecho para muchos investigadores como la partida de nacimiento de los archivos notariales.

Si bien el estudio arqueológico de los restos materiales conservados en cualquier naufragio antiguo nos permitía recuperar datos sobre las mercancías objeto de comercio, las rutas comerciales, los conocimientos náuticos y el sistema constructivo y, en el mejor de los casos, determinados aspectos relacionados con la vida a bordo, para momentos posteriores al siglo XIII la interpretación de este tipo de yacimientos se ve reforzada con la riquísima información recogida en los archivos históricos. Su análisis permite descubrir la vida del buque, desde su salida de la grada de construcción hasta su hundimiento, además de aspectos tales como los propietarios de los barcos, nacionalidad, nombre del piloto y de toda la tripulación, propietarios, características y valor de las mercancías, seguros marítimos que la cubrían y compañías que se dedicaban a ello, procedencia de la artillería, sistemas constructivos, además de aspectos tales como las costumbres y mentalidades de la gente del mar.

Una amplia y variada documentación repartida en multitud de archivos de diversa naturaleza (generales, municipales, militares, privados, etc.) que, por su origen (administrativo, judicial, religioso, privado, etc.), generó a lo largo de los siglos una gran diversidad de tipos y series documentales específicas que es preciso caracterizar y localizar como paso previo a cualquier proceso de estudio. Un fondo documental de millones de documentos que se encuentran ordenados y sistematizados según criterios archivísticos diferentes, provistos irregularmente de instrumentos de descripción (guía, índices, inventarios y catálogos) para facilitar el acceso y recuperación de los mismos y cuya base es el uso de términos normalizados (descriptores) de carácter, principalmente, onomástico, cronológico, geográfico, numérico y temático, este último, sin duda, el de mayor interés para los investigadores.

En función del nivel de descripción alcanzado para cada una de las secciones o series de archivos, y de la particularidad de la temática que se desea analizar, los términos normalizados de carácter temático pueden convertirse en un importante aliado del investigador, restringiendo considerablemente la ingente labor de búsqueda y lectura de documentos. En caso contrario, la búsqueda se hace más compleja, al tener que enfrentarse el investigador a la siempre ardua tarea de consultar miles de documentos que, en su mayor parte, no suelen ser de interés para la temática de estudio, con importantes consecuencias para el investigador en tiempo de dedicación y recursos.

Justificación del proyecto

Cuando se descubre un barco hundido, uno de los principales problemas para los arqueólogos consiste en identificarlo, determinando su nombre e historia. Partiendo de la información proporcionada por los escasos restos materiales hallados, es preciso rastrear en los archivos históricos aquellos datos que permitan su identificación, a partir de la nacionalidad, propietario, detalles de la artillería, sistema constructivo o el inventario de la carga que transportaba en su último viaje. Encontrar estos datos entre cientos de miles de documentos no indexados es una labor que consume gran cantidad de tiempo y recursos de investigación.

Los archivos guardan la historia de cientos de miles de naufragios históricos aún por localizar, datos que son la base para definir zonas de alto valor patrimonial y priorizar estrategias de investigación y/o protección. Junto a ello, la documentación cartográfica náutica, analizada desde una perspectiva espacial (ya sea diacrónica o sincrónica), es de gran utilidad para conocer cómo era el paisaje por el que navegaba cuando aconteció el naufragio, además de identificar la evolución que ha tenido a lo largo de los siglos, definiendo zonas de riesgo para la navegación (asociado a zonas de concentración de naufragios), nivel de aterramiento-erosión del lecho marino y, por ende, de enterramiento o exposición de los pecios (vinculado a la posibilidad de localizar pecios mediante técnicas de prospección superficial), etc. Líneas de investigación que consideramos imprescindibles para perfilar una correcta política de gestión de este patrimonio en relación con su investigación, protección y preservación y frente al posible impacto negativo asociado a procesos naturales o de obras en el litoral.

Desde el Área de Documentación del Centro de Arqueología Subacuática del Instituto Andaluz del Patrimonio Histórico (en adelante CAS-IAPH) se ha venido trabajando desde 1998 en esta línea, desarrollando un amplio proceso de investigación documental con cuatro objetivos bien definidos (Alonso/Márquez, 2013):

- Identificar las fuentes documentales relativas al patrimonio arqueológico subacuático andaluz, analizando su distribución geográfica y caracterizando las series documentales de mayor interés.
- Diseñar estrategias para la investigación y explotación de estas fuentes documentales.
- Sistematizar su información mediante el uso de herramientas de gestión normalizadas en cada uno de sus campos temáticos.
- Desarrollar proyectos propios de investigación para identificar el patrimonio cultural subacuático andaluz.

La labor directa de investigación realizada por el CAS-IAPH desde su creación ha sido amplia y productiva, si bien la limitación de los recursos disponibles para estos fines y la imposibilidad de realizar acuerdos de colaboración con grupos de investigación específicos, debido a la escasa implantación de la arqueología subacuática en el ámbito académico universitario, han condicionado considerablemente su desarrollo. Tras el proceso de identificación,

localización y caracterización de las instituciones y series documentales de mayor interés para el objeto de nuestro trabajo, se puso en marcha un amplio proyecto de investigación para la explotación de las mismas en diferentes fases y niveles. En la actualidad nuestro sistema de gestión de la información (SIGnauta) cuenta con cerca 1600 registros de naufragios para aguas andaluzas (Alonso *et alii*, 2007), superando los 377 inicialmente conocidos a través de diferentes trabajos de investigación (Chaunu, 1983; Serrano, 1991; García-Baquero, 1988; Lakey, 1987, entre otros).

Abordar una línea de estudios tan compleja y extensa como la de explotar la ingente documentación custodiada en los archivos históricos (nacionales e internacionales) es una labor que, con la metodología actual de estudio, solo es posible tras muchos años de esfuerzos por parte de un amplísimo equipo humano de trabajo, motivo por el que, desde 2012, el CAS-IAPH se planteó, en colaboración con el Centro de Reconocimiento de Formas y Tecnologías del Lenguaje (PRHLT) de la Universidad Politécnica de Valencia[1], diseñar un proyecto basado en el uso de las nuevas tecnologías para la indexación de términos no transcritos a partir de imágenes de documentos manuscritos, estrategia en la que nace el Proyecto Galeón que aquí presentamos.

Objetivos del proyecto

Los fondos de los archivos históricos conforman un importante legado cultural en los que se guardan aspectos aún no conocidos de nuestra historia política, social, económica y científico-tecnológica. La necesidad de conservar la herencia cultural escrita y hacerla accesible a la sociedad ha impulsado, en las últimas décadas, numerosos proyectos de digitalización y de mejora de la accesibilidad a través de buscadores y portales web especializados en el marco de una política de *Open Access*. Como resultado de ello, están disponibles para su visualización y estudio gran cantidad de imágenes digitales (cientos de terabytes de memoria de almacenamiento). Sin embargo, el verdadero problema, el de la accesibilidad selectiva a los contenidos textuales de dichas imágenes, está aún pendiente de resolver. Nos referimos a la posibilidad de buscar y encontrar entre millones de imágenes aquellas páginas que específicamente contienen los datos o informaciones que interesan para cada línea de investigación.

Con este fin nos planteamos desarrollar una propuesta teórica para buscar soluciones innovadoras, escalables y de bajo coste, que permitan realizar consultas documentales en grandes colecciones de imágenes de textos históricos manuscritos. Las herramientas a desarrollar deberían permitir para ello la localización de palabras previamente definidas entre una gran cantidad de legajos y documentos de series concretas. Como ámbito de aplicación se ha definido la posibilidad de trabajar sobre series especificas de los fondos del Archivo General de Indias (AGI), sin duda uno de los de mayor interés desde la perspectiva nacional e internacional para la arqueología subacuática.

El AGI conforma un depósito documental de gran valor que ilustra la historia del Imperio español de las Américas y las Islas Filipinas. Este archivo, declarado por la Unesco patrimonio de la humanidad en 1987, está compuesto por 43000 legajos, distribuidos a lo largo de ocho kilómetros de estanterías, con unos 80 millones de páginas. Entre sus series de mayor interés para la arqueología subacuática, están las de Indiferente de Justicia y Gobierno (1492-1870), generadas en base a los pleitos habidos entre armadores, aseguradoras y comerciantes en demanda de resarcimiento por pérdidas causadas por naufragio de la embarcación.

[1] https://www.prhlt.upv.es/ [Acceso 21-01-2015].

El AGI tiene buena parte de sus fondos digitalizados, pero en la actualidad la cantidad de páginas transcritas en comparación con la ingente masa documental histórica que custodia no es significativa. Debido a la dificultad para segmentar estos textos en caracteres tampoco es posible realizar búsquedas con técnicas tradicionales de reconocimiento óptico (OCR). Es obvio que la transcripción completa de los archivos es una tarea colosal, tan cara desde una vertiente económico-temporal que no es realista a medio plazo. Por otro lado, las técnicas que se utilizan actualmente para la transcripción automatizada o asistida aún no están suficientemente desarrolladas. Habrá que esperar años para comenzar a obtener resultados en este campo, por lo que la solución debía venir por otra vía.

Las palabras son elementos importantes del lenguaje que aportan, de manera directa, una información explícita acerca del asunto o temática sobre la que versan los documentos. Definir el número de términos o palabras de interés para efectuar búsquedas temáticas encaminadas, por ejemplo, a la localización de información sobre naufragios históricos, es una labor bien definida y abarcable a corto plazo, pero incluir los mismos en los instrumentos de descripción al uso en cada archivo se convierte ya en un proyecto difícilmente ejecutable a medio o largo plazo, máxime si se hace extensible a muchos otros campos de investigación, más allá de los de interés exclusivo para la arqueología subacuática.

Los avances que a nivel de las nuevas tecnologías se están produciendo en los últimos años de la mano del reconocimiento automatizado de caracteres impresos y manuscritos en soporte digital y de la traducción automatizada, introducen nuevas perspectivas de desarrollo en este campo que es necesario explorar. Hoy día es posible indexar imágenes de documentos manuscritos de forma que se pueda realizar sobre las mismas búsquedas de texto libre. En este tipo de búsquedas es difícil predecir de antemano cuáles son los términos o palabras claves que van a ser útiles al usuario de la herramienta para localizar la información deseada. Lo ideal sería permitir al investigador o usuario una búsqueda libre, es decir, utilizar en dicha operación cualquier palabra. Sin embargo, como comentaremos más adelante, por motivos de eficacia la mejor opción es partir de un lenguaje normalizado previamente. Es decir, de un glosario de palabras claves o descriptores definidos previamente para un campo específico de estudio, en nuestro caso, términos vinculados al naufragio o a la recuperación de buques históricos.

Metodología de trabajo. Situación actual

Hasta el momento actual, la búsqueda de información textual sobre cientos de miles o millones de imágenes de manuscritos históricos se suele considerar inviable, pues las tecnologías prevalentes para esta tarea conllevan costes prohibitivos en términos computacionales y/o de recursos humanos. Por ello, se suele abordar a través, principalmente, de dos tipos de estrategias:

- La primera consiste en usar técnicas de reconocimiento óptico (OCR) sobre texto electrónico. Pero para esto el paso previo es convertir las imágenes en texto; es decir, hay que realizar un proceso previo de transcripción manual o semiautomática.
- La segunda se basa en técnicas conocidas en inglés como *Key Word Spotting* (KWS), a las que nos referiremos por su traducción directa como «Búsqueda de Palabras Clave» (BPC).

El punto débil de la primera aproximación es el coste prohibitivo en recursos humanos necesarios. Esto es obvio si la transcripción se lleva a cabo de forma puramente manual (como suele ser el caso generalmente). Pero aún en el caso de usar la tecnología del Reconocimiento de Texto Manuscrito (RTM), estas no están lo suficientemente desarrolladas para producir resultados utilizables en el campo de la investigación documental histórica. Los productos comerciales de transcripción están basados en la tecnología de reconocimiento de caracteres aislados

(OCR en inglés) desarrollada durante las dos últimas décadas. Esta tecnología requiere una segmentación de las imágenes a nivel de carácter, que es prácticamente imposible aplicar en la documentación textual manuscrita.

Por otro lado, las aproximaciones holísticas modernas al RTM (Bazzi/Schwartz/Makhoul, 1999; Marti/Bunke, 2001: 65-90; Toselli *et alii*, 2004: 519-539) no necesitan ningún tipo de segmentación a nivel de carácter o palabra. En el actual estado de desarrollo, los prototipos del RTM pueden alcanzar niveles de precisión que oscilan entre el 60% y el 90% de palabras bien reconocidas (Marti/Bunke, 2001: 65-90; Toselli *et alii*, 2004: 519-539) para textos manuscritos de gran calidad. Sin embargo la precisión suele degradarse muy significativamente para textos históricos. En las figuras 1, 2 y 3 se muestran algunas imágenes de textos de interés para el Proyecto Galeón. En general la precisión esperada es demasiado baja para ser usada directamente en sistemas de búsqueda de texto convencionales. Además, dada la elevada tasa de error, el coste humano de corregir los errores también es prohibitivo para los grandes volúmenes de imágenes involucrados. Lamentablemente, esto sigue siendo cierto incluso si se utilizan técnicas recientes de transcripción interactiva con las que el esfuerzo humano se puede reducir notablemente (Romero/Toselli/Vidal, 2012).

Una alternativa reciente que puede abaratar considerablemente los costes humanos de transcripción es el planteamiento de colaboración conocido como *Crowdsourcing*. En este mar-

Figura 1. Los protocolos notariales son una fuente de gran valor para localizar información sobre testamentos de náufragos o contratos de compraventas de embarcaciones y mercancías (Fondo gráfico CAS-IAPH. Foto: L. Márquez).

Figura 2. El Archivo General de Indias es el organismo en el que se conserva un mayor volumen de información sobre naufragios históricos.

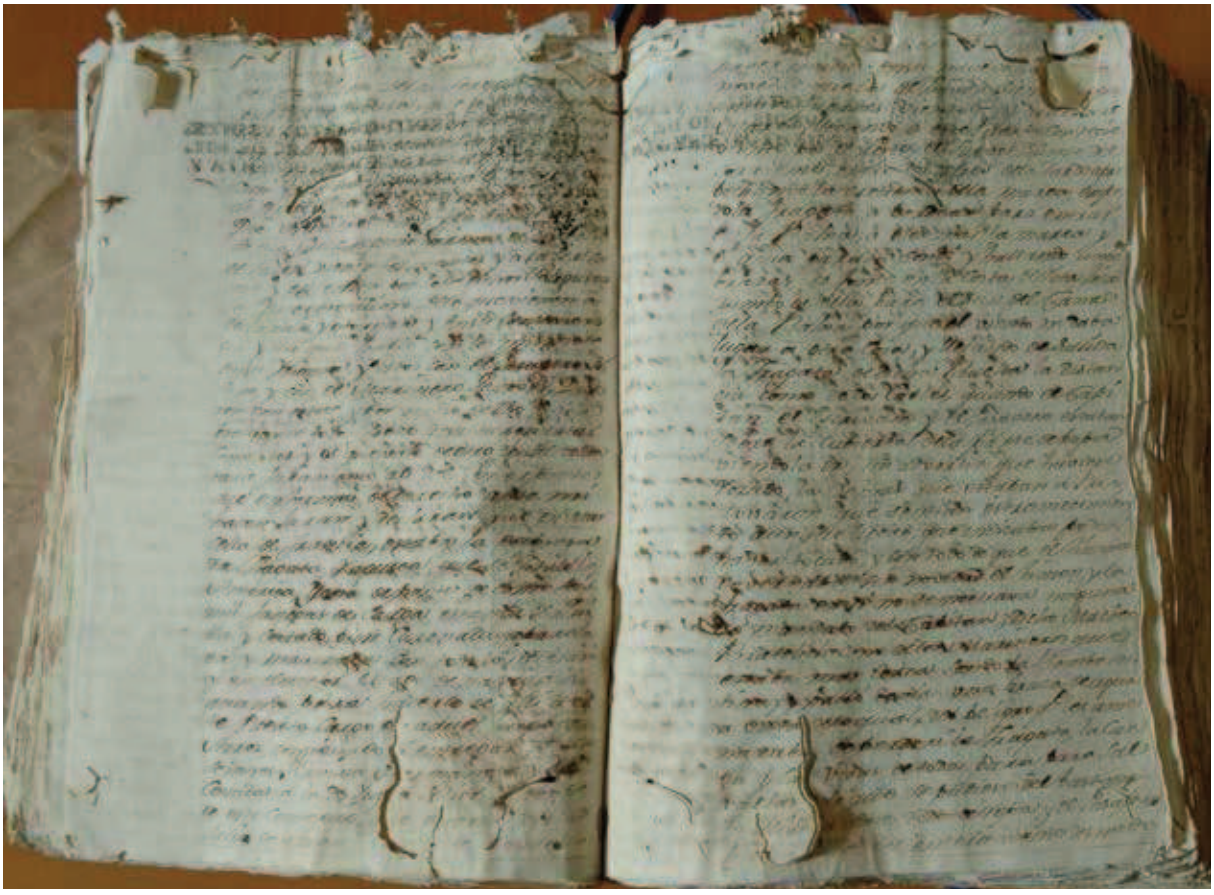

Figura 3. El mal estado de conservación de los documentos es uno de los problemas que se plantean para la aplicación de reconocimiento formal de caracteres manuscritos.

co, una gran cantidad de personas, no necesariamente expertas, colaboran de manera voluntaria y gratuita, anotando, editando, corrigiendo, transcribiendo o traduciendo grandes colecciones. Este sistema ha logrado notables éxitos en otros campos, como por ejemplo en la comunidad Galaxy-Zoo[2], que ha contribuido a clasificar millones de galaxias, o en el programa de digitalización de prensa llevado a cabo por la Biblioteca Nacional de Australia[3] donde se han corregido centenares de miles de páginas producidas por sistemas de OCR. Recientemente se ha utilizado en proyectos de transcripción de textos manuscritos tales como «Old Weather»[4] y «Ancestry.com's World Archives»[5] donde el material se presentaba en forma de formularios, segmentado o al menos de manera fácil de descifrar y comprender.

Un paso más allá lo dan los proyectos «Transcribe Bentham»[6] y «TranScriptorium»[7]. En el primero de ellos se realiza la transcripción manual directa de los manuscritos del famoso filósofo y reformador inglés del siglo XVIII y XIX Jeremy Bentham. Mientras que en TranScriptorium, los voluntarios utilizan un prototipo de asistencia predictiva de tal manera que interaccionan con el sistema automático guiándolo para encontrar la transcripción perfecta de los textos (Romero/Toselli/Vidal, 2012).

Lamentablemente, el sistema de *Crowdsourcing* no es directamente utilizable en los documentos de interés para la arqueología subacuática pues el volumen de imágenes de interés que hay que transcribir es muchísimo mayor y de mucha mayor dificultad que en los casos comentados.

La segunda aproximación arriba indicada, de «Búsqueda de Palabras Clave» BPC, consiste en localizar las posiciones en documentos o colecciones donde es probable que una (o varias) palabra(s) dada(s) aparezcan. Con esta técnica, el grado mínimo de probabilidad, o «confianza», de cada búsqueda lo especifica el usuario de forma más o menos explícita. En BPC podemos encontrar dos planteamientos, llamados *query-by-example* y *query-by-string*. El primero de ellos necesita una imagen de ejemplo para encontrar imágenes similares.

Encontrar imágenes adecuadas de una palabra de interés en los citados documentos es una tarea ardua. Generalmente, para localizar un ejemplo del término que se busca hay que examinar una ingente documentación hasta encontrar una muestra suficientemente representativa. Además, la grafía de las palabras puede ser variable en función de la cronología. Incluso para un mismo periodo de tiempo, se observa una gran variedad de estilos, dependiendo del escribano; e incluso para un mismo escritor, dependiendo del momento y posición en que escribió, si el texto era original o copia, etc. Por otra parte, por motivos computacionales la escalabilidad de las técnicas de *query-by-example* es muy limitada, no permitiendo obtener ventaja del contexto textual en el que está escrita la palabra que se busca para aumentar la precisión de la búsqueda.

El planteamiento de BPC mediante *query-by-string* es más reciente. En este caso, el usuario especifica el término que busca como una cadena de caracteres (o *string*) (Frinken/Fischer/Manmatha, 2010: 352-357; Khurshid/Faure/Vincent, 2012: 2598-2609; Marinai/Marino/Soda, 2006: 1187-1199; Rodríguez-Serrano/Perronnin, 2009: 351-355; Vamvakas *et alii*, 2008: 525-532). La mayoría de estas técnicas requieren una presegmentación a nivel de carácter o de palabra, obteniéndose buenos resultados principalmente en documentos impresos. En el caso de textos manuscritos, solo unas pocas aproximaciones (Rodríguez-Serrano/Perronnin, 2009: 351-355;

---

[2] http://www.galaxyzoo.org [Acceso 21-01-2015].
[3] http://trove.nla.gov.au/system/stats#corrections [Acceso 21-01-2015].
[4] http://www.oldweather.org [Acceso 21-01-2015].
[5] http://landing.ancestry.com/wap/learnmore.aspx [Acceso 21-01-2015].
[6] https://www.ucl.ac.uk/Bentham-Project [Acceso 21-01-2015].
[7] http://www.transcriptorium.eu [Acceso 21-01-2015].

Frinken/Fischer/Manmatha, 2010: 352-357) han sido probadas con un éxito relativo. La mayoría de estas técnicas se basan más o menos explícitamente en información derivada de algún proceso de RTM. Entre los trabajos recientes que siguen esta tendencia hay que reseñar los siguientes (Fischer *el alii,* 2012: 934-942; Frinken *et alii,* 2012: 211-224; Toselli *et alii*, 2013). De ellos, las mejores prestaciones de BPC se obtienen con el método propuesto por Frinken y otros (Frinken *et alii,* 2012: 211-224), basado en redes neuronales recurrentes.

Este método resulta inviable en la práctica, debido al exorbitante coste computacional de su fase de entrenamiento. Además, el método que se propone en Fischer y otros (2012: 934-942) se basa en un modelado morfológico de los caracteres mediante modelos ocultos de Markov (en inglés *Hidden Markov Models*, HMM) de caracteres. Aunque sus prestaciones de BPC son inferiores, su coste de entrenamiento es muchísimo más bajo. No obstante la búsqueda «sobre la marcha» inherente a este método, resulta prohibitiva, incluso para colecciones relativamente pequeñas. En un trabajo posterior (Toselli/Vidal, 2013: 501-505), Toselli y Vidal reducen este coste drásticamente gracias a una representación de las imágenes de texto mediante grafos de caracteres, manteniendo las prestaciones de BPC originales.

Como conclusión podemos decir que las aproximaciones tradicionales BPC requieren altos costes computacionales debido, sobre todo, al proceso de entrenamiento necesario, o bien, a la necesidad de explorar todo el espacio de búsqueda en tiempo de consulta. En cualquiera de los dos casos, la puesta a punto del sistema sería prohibitivamente lenta.

Como opción, nos planteamos que una aplicación de estas características debería fundamentarse en las técnicas introducidas por Toselli, Vidal y otros (en Toselli *et alii*, 2013), basadas en Grafos de Palabras (GP) derivados de RTM. Estas técnicas tienen la ventaja de facilitar el uso de un modelo de lenguaje durante el proceso de reconocimiento, gracias al cual se dispone de valiosa información derivada del contexto léxico, lo que permite realizar búsquedas mucho más precisas.

Objetivos del proyecto

Como se acaba de mencionar, nuestro principal objetivo va encaminado a diseñar un modelo teórico de trabajo para desarrollar y adaptar la tecnología de Búsqueda de Palabras Claves (BPC) basada en Grafos de Palabras (Toselli *et alii*, 2013) a la localización de información en documentación histórica de carácter textual manuscrita relacionada con el patrimonio cultural subacuático (si bien aplicable a cualquier otro campo del saber) en fondos de archivos. Para ello apostamos por un modelo que, partiendo de la indexación de un conjunto de términos suficientemente amplio para permitir trabajar sobre grandes volúmenes documentales, sea capaz de realizar búsquedas probabilísticas y recuperar la información jerárquica no convencional de manera rápida con diferentes niveles de granularidad (línea, página, legajo, etc.) y con un parámetro de confianza que ayude a definir el equilibrio deseado entre precisión y cobertura.

Resultados esperados

El modelo propuesto de BPC puede desarrollarse según medidas probabilísticas sobre la confianza de que una palabra aparezca (o no) en una línea, página, documento, legajo o colección. Usando estructuras de datos adecuadas (construidas en una fase preparatoria) que soporten los índices y la correspondiente información probabilística, la carga computacional necesaria para cada búsqueda puede ser insignificante y prácticamente independiente del número de imágenes de las colecciones consideradas. Incluso para inmensos volúmenes de documentos, los tiempos de búsqueda esperados pueden ser de alrededor de unos pocos segundos.

Con Galeón queremos estudiar el alcance de estas técnicas, así como su escalabilidad y viabilidad para ser usadas en aplicaciones a gran escala. Una técnica que permitiría, en un futuro:

- Realizar búsquedas de manera automatizada sobre grandes volúmenes de imágenes de texto manuscrito sin necesidad de transcribirlas previamente.
- Obtener resultados empíricos sobre la viabilidad y escalabilidad del uso de estas técnicas en tareas de indexación en colecciones de gran volumen documental.
- Orientar la planificación de futuras propuestas de proyectos de indexación y búsqueda de gran envergadura.
- Generar índices jerárquicos para conjuntos representativos de imágenes de texto manuscrito de series de interés para la investigación del patrimonio arqueológico subacuático.
- Y, sobre todo, reducir considerablemente el coste de la fase de investigación documental necesaria para, en este caso, el estudio del gran numero de naufragios que reposan en las aguas de España, Portugal y América, si bien, es evidente que este modelo podría ser aplicado a cualquier otro fondo documental manuscrito (independientemente de la temática de investigación), previa definición de las correspondientes palabras claves y del imprescindible proceso de enseñanza-aprendizaje del sistema.

Sin duda una base firme para futuras acciones de transferencia tecnológica a la industria o para otros posibles modelos de explotación.

Bibliografía

Alonso Villalobos, C., y Márquez Carmona, L. (2013): «Fuentes de información del patrimonio arqueológico subacuático de Andalucía: una década de investigación documental». En *I Congreso de Arqueología Náutica y Subacuática Española. Cartagena, 14, 15 y 16 de marzo de 2013*. Nieto Prieto, X.; Ramírez Pernía, A., y Recio Sánchez, P. (coords.). Madrid, Ministerio de Educación, Cultura y Deporte, pp. 751-763.

Alonso Villalobos, C. *et alii* (2007): «SIGNauta: un sistema para la información y gestión del patrimonio arqueológico subacuático de Andalucía». En *Boletín del Instituto Andaluz del Patrimonio Histórico,* n.º 63, pp. 26-41.

Bazzi, I.; Schwartz, R., y Makhoul, J. (1999): «An Omnifont Open-Vocabulary OCR System for English and Arabic». En *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n.º 6, pp. 495-504.

Bosch, V.; Toselli, A. H., y Vidal, E. (2012): «Statistical text line analysis in handwritten documents». *Proceedings of the 13th International Conference on Frontiers*. En *Handwriting Recognition. ICFHR 2012, 18-20 September 2012, Bari, Italy*. Los Alamitos, Calif., IEEE Computer Society, pp. 201-206.

Chaunu, P. (1983): *Sevilla y América: siglos xvi y xvii*. Sevilla, Universidad de Sevilla.

Frinken, V. *et alii* (2010): «Adapting BLSTM neural network based keyword spotting trained on modern data to historical documents». En *Proceedings Of the 12th International Conference on Frontiers in Handwriting Recognition. ICFHR 2010, Kolkata, India, 16-18 November 2010*, Los Alamitos, Calif., IEEE Computer Society, pp. 352-357.

— (2010): «Lexicon-free handwritten word spotting using character HMMs». En *Pattern Recognition Letters*, vol. 33, n.º 7, pp. 934-942. [Special issue on Awards from International Conference on Pattern Recognition].

— (2012): «A Novel Word Spotting Method Based on Recurrent Neural Networks». En *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, n.º 2, pp. 211-224.

García-Baquero González, A. (1988): *Cádiz y el Atlántico (1717-1778): el comercio colonial español bajo el monopolio gaditano*. Cádiz, Diputación de Cádiz.

Khurshid, K.; Faure, C., y Vincent, N. (2012): «Word spotting in historical printed documents using shape and sequence comparisons». En *Pattern Recognition*, vol. 45, n.º 7, pp. 2598-2609.

Lakey, D. C. (1987): *Shipwrecks in the Gulf of Cadiz: A Catalog of historically documented wrecks from the fifteenth through the nineteenth centuries. Final Report*. Institute of Nautical Archaeology (INA).

Marti, U. V., y Bunke, D. H. (2001): «Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition System». En *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, n.º 1, pp. 65-90.

Marinai, S.; Marino, E., y Soda, G. (2006): «Font adaptive word indexing of modern printed documents». En *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, n.º 8, pp. 1187-1199.

Rodriguez-Serrano, J. A., y Perronnin, F. (2009): «Handwritten word image retrieval with synthesized typed queries». En *Proceedings of the 10th International Conference on Document Analysis and Recognition. ICDAR'09, 26-29 July, Barcelona, Catalonia, Spain*. Los Alamitos, Calif., IEEE Computer Society, pp. 351-355.

Romero, V.; Toselli, A. H., y Vidal, E. (2012): *Multimodal Interactive Handwritten Text Transcription*. Singapore, World Scientic Publishing. (Series in Machine Perception and Articial Intelligence, vol. 80).

Serrano Mangas, F. (1991): *Naufragios y rescates en el tráfico indiano durante el siglo xvii*. Madrid, Sociedad Estatal para el Quinto Centenario.

Toselli, A. H., y Vidal, E. (2013): «Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents». En *Proceedings of the 12th International Conference on Document Analysis and Recognition. ICDAR '13, 25-28 August 2013 Washington, D.C.* Piscataway, NJ, IEEE Computer Society, pp. 501-505.

Toselli, A. H. *et. alii* (2004): «Integrated Handwriting Recognition and Interpretation using Finite- State Models». En *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, n.º 4, pp. 519-539.

Toselli, A. H. *et alii* (2013): *Word-graph based keyword spotting and indexing of handwritten document images. Technical report*. Universidad Politécnica de Valencia.

Vamvakas, G. *et alii* (2008): «A complete optical character recognition methodology for historical documents». En *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems. DAS '08, September 16-19, 2008, Nara, Japan*. Los Alamitos, Calif., IEEE Computer Society, pp. 525-532.