Transiciones en la agricultura
y la sociedad rural
II Congreso Internacional
XVI SEHA | VII RuralReport
TransRuralHistory Compostela 2018

El reconocimiento automático de texto manuscrito aplicado a la Contaduría de Hipotecas de Girona

Vicente Bosch (UPV), Rosa Congost (UdG), Lorenzo Quirós (UPV), Enric Saguer (UdG) y Enrique Vidal (UPV)

PANEL ID: S140: Nuevas fuentes y nuevas metodologías aplicadas a la investigación en historia agraria

---

Abstract:

Los historiadores llevan tiempo accediendo a repositorios de textos impresos cuyos caracteres, capturados como imagen, han sido decodificados mediante programas de reconocimiento óptico (OCR) y permiten la realización de búsquedas básicas y otras operaciones analíticas. Sin embargo, los textos manuscritos se han resistido hasta el momento a este tipo de tratamientos, entre otros motivos por la variabilidad de la escritura a mano y por la dificultad que presenta segmentar cada uno de los caracteres que integran una palabra. El desarrollo reciente de técnicas de reconocimento de texto manuscrito (RTM) permite albergar nuevas expectativas sobre la posibilidad de tratar masivamente grandes conjuntos documentales. El objetivo de esta comunicación es dar cuenta de los resultados obtenidos en un proyecto de colaboración entre el *Centre de Recerca d'Història Rural* de la Universitat de Girona y el grupo *Pattern Recognition and Human Language Technology* de la Universidad Politécnica de Valencia centrado en la transcripción asistida de los libros de la Contaduría de Hipotecas de Girona y el etiquetado (xml) automático de las tipologías documentales, los antropónimos, los topónimos y los oficios.

Keywords: Registro de Hipotecas, reconocimiento de texto manuscrito

---

S E H A

Rural RePort
Rede de História Rural em Português

histagra
Grupo de Historia Agraria e Política do Mundo Rural
USC

Los progresos realizados en las dos últimas décadas en el ámbito de las tecnologías de reconocimiento de texto han sido muy importantes y han permitido poner a disposición de los investigadores cantidades ingentes de documentación impresa (boletines, periódicos, libros, folletos…) generalmente dispersa y poco accesible. Junto con el desarrollo de herramientas de búsqueda, también han abierto la posibilidad de gestionar selectivamente y de forma manejable este gran volumen información, anteriormente sólo disponible como imagen. Sin embargo, dichas tecnologías, conocidas habitualmente bajo las siglas OCR (Optical Character Recognition), exitosas en la decodificación de signos tipográficos individualizables, han fracasado en su aplicación a la documentación manuscrita, básicamente por el carácter irregular, continuo y ligado de la escritura. No ha sido hasta momentos más recientes que se han empezado a poner en marcha técnicas de reconocimento de texto manuscrito cuyos resultados resultan esperanzadores desde la perspectiva de su potencial utilización para la transcripción de documentación histórica. Actualmente estas técnicas no sustituyen completamente la intervención humana, pero permiten procesos semiautomáticos de transcripción, con reducciones incrementales del esfuerzo a medida que el sistema informático aprende de las decisiones expertas del investigador.[1]

El objetivo de esta comunicación es dar a conocer el enfoque y los primeros resultados de un proyecto de colaboración en este ámbito entre el *Centre de Recerca d'Història Rural* de la Universitat de Girona y el *Pattern Recognition and Human Language Technology Research Center* (PRHLT) de la Universidad Politécnica de Valencia. Venimos trabajando en él de forma tentativa y preliminar desde junio de 2016, pero a partir del pasado diciembre de 2017, ha adquirido mayor envite al conseguir financiación a través del programa de *Ayudas a Equipos de Investigación Científica* de la Fundación BBVBA, en su convocatoria de Humanidades Digitales.[2] Concretamente, esta colaboración busca aplicar las herramientas y procesos de reconocimiento de texto manuscrito a las imágenes digitalizadas de una gran serie documental: los libros de las Oficinas de Hipotecas (1768–1861) de la región de Girona.

La colaboración de ambos equipos en el tratamiento de esta serie constituye un punto de encuentro entre dos líneas de trabajo con objetivos distintos, de desarrollo tecnológico en un caso, de análisis del cambio social en el otro. Desde la perspectiva de la investigación histórica, la oportunidad de digitalizar y transcribir una serie documental de larga duración y una cobertura territorial bastante amplia, abre un escenario nuevo para el análisis de las dinámicas sociales, precisamente durante un periodo de grandes transformaciones. La posibilidad de disponer de centenares de miles de registros referidos, como se verá, a transacciones de índole diversa, nos permitirá nuevos acercamientos al análisis de los procesos de cambio

[1] Algunos ejemplos en Romero et al. (2013) y en Toselli et al. (2017).

[2] El proyecto se titula *Explorando cambios sociales silenciosos: una propuesta a partir de la explotación digital de una gran mina de datos históricos (Cataluña, siglo XVIII)*. Los primeros pasos de esta investigación se han realizado en el marco del proyecto HAR2014-54891-P *Ni élites ni pobres. Las clases medias en perspectiva histórica*, financiado por el Ministerio de Economía y Competitividad.

social, fijándonos especialmente en aquellos que son más silenciosos porque afectan a grupos sociales menos presentes en los registros documentales. Este es nuestro reto.

En las páginas que siguen revisaremos, en primer lugar, el estado actual de las tecnologías de reconocimiento de texto manuscrito; continuaremos exponiendo las líneas centrales de nuestro proyecto; describiremos, en tercer lugar, las características de la fuente documental y sus posibilidades de tratamiento masivo y sistemático; y, finalmente, detallaremos los resultados obtenidos hasta el momento en el proceso de transcripción asistida.

Las tecnologías de reconocimiento de texto

El desarrollo de tecnologías capaces de reconocer e interpretar automáticamente texto escrito ha tenido desde hace varias décadas un gran interés académico, social e industrial. Esto es particularmente cierto cuando el texto está manuscrito en papel o en algún otro soporte.

Ante la aparición de las nuevas tecnologías informáticas, el interés por el reconocimiento automático de texto manuscrito decayó durante algún tiempo, bajo la asunción de que estas tecnologías pronto harían desaparecer los documentos de papel y con ellos la necesidad de procesar texto manuscrito. Sin embargo, más recientemente, el reconocimiento de documentos de texto manuscrito ha vuelto a ser un tema candente de investigación y desarrollo, al constatar la ingente cantidad de manuscritos históricos que se conservan en archivos y bibliotecas de todo el mundo, muchos de los cuales se han estado digitalizando las últimas décadas para dejarlos al alcance del público en general.

No es de extrañar, por tanto, el gran interés social y comercial en hacer posible el acceso simple y ágil al inmenso legado de información histórica, política, económica, demográfica, y cultural en general, contenido en dichos textos.

Sin embargo, para que las imágenes de texto manuscrito sean realmente útiles deben ser anotadas con información acerca de su contenido. Lógicamente la información más rica acerca del contenido de una imagen de texto es precisamente su transcripción. Dado el inmenso volumen de documentos de interés involucrados, dicha transcripción no puede obtenerse de manera manual, por lo que el proceso pasa necesariamente por el uso de métodos automatizados o semi-automatizados.

En el caso de documentos impresos, la transcripción automática se ha venido abordando desde hace algunas décadas con técnicas de reconocimiento óptico de caracteres (más conocidas como OCR por su expresión en inglés “Optical Character Recognition”). Los resultados obtenidos mediante estas técnicas son bastante variables y dependientes de la calidad de los documentos. Pero para documentos en buenas condiciones, las tasas de acierto de caracteres pueden estar por encima de 99% (lo que significa alrededor de un 5% de palabras con algún error). Es importante destacar que las herramientas de OCR actuales se basan en una segmentación explícita de los caracteres que aparecen en el documento digitalizado. Si la segmentación en caracteres (o incluso palabras) no funciona correctamente

entonces la precisión de reconocimiento se reduce drásticamente, hasta hacer prácticamente inservible el resultado.

Cuando los documentos digitalizados son textos manuscritos, la segmentación explícita de los caracteres es simplemente imposible y las técnicas de OCR no son de ninguna utilidad. En la mayoría de documentos manuscritos históricos, ni tan sólo la segmentación en palabras es posible, y es necesario recurrir a técnicas holísticas que no requieren segmentación previa en palabras ni en caracteres y reconocen de forma integrada líneas de texto completas.

Estas técnicas, que con frecuencia se denominan simplemente "Reconocimiento de Texto Manuscrito" (RTM) (en inglés "Handwritten Text Recognition" -HTR-), utilizan en la actualidad conceptos de Reconocimiento de Formas, Aprendizaje Automático y Lingüística Computacional, tales como Modelos de Markov, Modelos de Lenguaje, Redes Neuronales y Aprendizaje Profundo.

En los últimos años ha habido notables avances en el campo de RTM. Sin embargo, las transcripciones que se obtienen con estos sistemas aún están lejos de ser perfectas.

Dada la complejidad de la mayoría de los documentos manuscritos de interés, para obtener transcripciones sin (demasiados) errores es preciso revisar las transcripciones producidas por estos sistemas. Trabajos recientes han estudiado técnicas interactivas que integran al usuario en el proceso de RTM (Romero, Toselli, & Vidal, 2012), obteniendo resultados sumamente alentadores.

Por otra parte, es importante destacar que con la tecnología de RTM actualmente disponible es posible desarrollar sistemas de indexación y búsqueda de contenidos en imágenes de documentos manuscritos. Estas técnicas de localización de términos o *palabras clave*, o *key word spotting* (KWS) en inglés, permiten explorar colecciones de imágenes de texto sin transcribir para encontrar aquellas imágenes en las que una determinada palabra o frase puede aparecer con un grado de confianza dado (Bluche et al., 2017; Fischer, Keller, Frinken, & Bunke, 2012; Frinken, Fischer, Manmatha, & Bunke, 2012).

Reconocimiento de texto manuscrito

El RTM es una tarea de gran desafío en el reconocimiento de formas. Aunque el texto está básicamente compuesto de caracteres, las aproximaciones tradicionales de reconocimiento de caracteres aislados, tal y como hemos comentado anteriormente, generalmente fracasan en la tarea de RTM. Esto se debe a la imposibilidad material de segmentar de manera fiable un texto continuo en sus caracteres individuales. Sin embargo, los seres humanos realizan estas tareas de segmentación y reconocimiento de una manera natural y sin aparente esfuerzo. La precisión se alcanza gracias a una fuerte inter-cooperación entre diferentes niveles de conocimiento: visual, morfológico, léxico, sintáctico y semántico. En este campo, las técnicas existentes de mayor éxito están basadas en la inter-cooperación de las mencionadas fuentes de conocimiento para conseguir un reconocimiento global.

Los primeros desarrollos en RTM aparecieron hacia finales de los años 60 con aplicaciones restringidas que implicaban vocabularios limitados, tales como el reconocimiento de direcciones postales o cheques bancarios. Sin embargo, no fue hasta varias décadas después cuando estos desarrollos recibieron un fuerte impulso, gracias al uso de tecnologías heredadas del Reconocimiento Automático del Habla (Kim, Govindaraju, & Srihari, 1999; Makhoul, Schwartz, Lapre, & Bazzi, 1998; Plamondon & Srihari, 2000; Steinherz, Rivlin, & Intrator, 1999), como los bien conocidos Modelos de Lenguaje (N-gramas) o los modelos ocultos de Markov (Jelinek, 1998). Recientemente, esta tecnología ha obtenido una considerable mejora al introducirse el uso de las redes neuronales (Bluche, 2015; Graves et al., 2009) para el modelado morfológico de los caracteres, obteniéndose tasas de error cercanas al 5% al nivel de caracteres (Dempster, Laird, & Rubin, 1977; Sánchez, Toselli, Romero, & Vidal, 2015). Un aspecto muy destacable de estos modelos estadísticos es que pueden aprenderse automáticamente a partir de ejemplos (Dempster et al., 1977). Esto es, dado un conjunto de imágenes de líneas, párrafos o páginas con su correspondiente transcripción (no necesariamente alineada a nivel de carácter ni de palabra), existen algoritmos robustos que estiman automáticamente los parámetros de estos modelos. Posteriormente, estos modelos pueden utilizarse para transcribir imágenes que no han sido vistas con anterioridad. Por tanto, esta tecnología se pueda aplicar fácilmente a cualquier idioma o sistema de escritura, lo que reduce notablemente los costes de desarrollo de sistemas de RTM.

El RTM se divide en dos etapas fundamentales. La primera etapa consiste en el análisis de la imagen de un documento con el fin de localizar y extraer las diferentes partes a transcribir. En la actualidad la unidad de transcripción más pequeña con la que se trabaja suele ser la línea, aunque idealmente se debería trabajar con unidades de mayor tamaño (párrafos o páginas) puesto que el contexto[3] en el proceso de transcripción es sumamente importante. Para esta etapa se utilizan técnicas de Análisis de Imágenes de Documentos (en inglés Document Layout Analysis) (Fiel, Grüning, Gatos, Diem, & Kleber, 2017; Pastor i Gadea, 2007), las cuales se encargan de limpiar la imagen, normalizar su geometría, detectar las zonas o bloques de texto, y detectar y extraer las líneas (o unidades mínimas de transcripción). En la segunda etapa, se realiza el propio proceso de transcripción, mencionado en el párrafo anterior.

Transcripción asistida de texto manuscrito

Aunque el RTM ha avanzado notablemente en la última década hasta obtener prototipos con resultados muy satisfactorios (precisión en algunos casos por encima del 80% de palabras correctas), los resultados obtenidos están lejos de ser transcripciones perfectas. Estos sistemas son de gran utilidad en tareas restringidas con vocabularios pequeños y con restricciones en el estilo de escritura. Sin embargo, en tareas reales sin ningún tipo de restricción, la tecnología actual proporciona resultados erróneos.

Si bien es cierto que la tecnología RTM disponible actualmente puede ser útil para indexar y realizar búsquedas en documentos, si lo que realmente se desea es obtener una transcripción de calidad o sin errores, es necesario que un experto humano (típicamente un paleógrafo en el

[3] Por contexto se entiende aquí las palabras cercanas a la palabra que se está transcribiendo.

caso de documentos antiguos) realice un trabajo de revisión y corrección. El escenario más usual que se contempla para este proceso consiste en realizar dicha corrección después de obtener los resultados de reconocimiento; esto es, como un proceso de post-edición. Sin embargo, la post-edición puede resultar ineficiente y poco cómoda, además de económicamente cara (téngase en cuenta que para este proceso se suelen necesitar expertos en paleografía). Una solución alternativa es utilizar técnicas interactivo-predictivas (Toselli, Romero, Pastor i Gadea, & Vidal, 2010; Toselli, Vidal, & Casacuberta, 2011).

La aproximación interactivo-predictiva supone un marco más cómodo y eficiente para el transcriptor humano, donde se combina la eficiencia de los sistemas de RTM con la precisión de los expertos, permitiendo una transcripción perfecta de las imágenes con un coste admisible. El proceso de interacción se basa en el *feedback* proporcionado por el usuario. Durante el proceso de transcripción, el sistema tiene en cuenta tanto la imagen de texto que se está transcribiendo como una porción de la transcripción validada por el usuario. La nueva salida proporcionada por el sistema es una nueva solución óptima que tiene en cuenta ambas informaciones. Este bucle continúa hasta que obtener una transcripción satisfactoria. La tecnología interactivo-predictiva actual se basa en lo que se conoce como grafo de palabras (Romero et al., 2012).

Cabe destacar que, además de ser un marco más cómodo para el transcriptor, la transcripción asistida puede ahorrar trabajo al corrector humano en el proceso de corrección de errores con respecto a utilizar la corrección tradicional de post-edición (Toselli et al., 2017). Esta opción es la que estamos aplicando al Registro de Hipotecas de Girona y a la que se referirán los resultados relatados en el último apartado.

El proyecto

Desde una perspectiva exclusivamente analítica y vinculada a la historia rural y social, el principal interés del proyecto es disponer de un gran banco de datos de carácter socioeconómico que permita analizar procesos de cambio social desde mediados del siglo XVIII. Tenemos especial interés en profundizar sobre los procesos menos detectables, los *cambios sociales silenciosos*, como la emergencia de los *menestrals*, un grupo social que, procedente del colectivo de trabajadores rurales, consiguió consolidar un cierto nivel de riqueza que le permitió distinguirse, llegando a adoptar una etiqueta social diferenciada (Congost, 2014; Congost, Ros, & Saguer, 2016). También nos interesa particularmente detectar la actividad de los trabajadores en el mercado de la tierra y del crédito. Los datos recabados y su análisis nos proporcionarán una base sólida para revisar, en un marco territorial supracomarcal, cuestiones históricamente relevantes, como la existencia (o no) de procesos de proletarización a fines del Antiguo Régimen, las condiciones de vida más humildes, la existencia de procesos de empobrecimiento y enriquecimiento, o la emergencia de nuevos grupos medios. Creemos que el banco de datos que obtendremos nos permitirá analizar, de forma dinámica y desde una perspectiva temporal de cierto alcance, numerosas trayectorias individuales y colectivas, relativas a aquellos grupos sociales sobre los cuales conocemos pocos detalles.

El marco territorial de análisis va a ser la región de Girona. El Registro de Hipotecas del distrito de Girona es una fuente rica y sistemática, aunque como expondremos más adelante, su ámbito territorial experimentó sucesivas divisiones. En el momento de su creación, el registro de Girona cubría un territorio de 3.884 km^2; posteriormente, a partir de 1774, se crearon nuevos Oficios que lo redujeron a unos 1.422 km^2, y a partir de 1829 volvió a reducirse otra vez. Aunque la existencia del Registro de Hipotecas comprende desde 1768 hasta 1862, en la primera fase del proyecto nos planteamos cubrir solamente el período 1768-1805. La razón de esta limitación inicial –que no excluye que en fases posteriores se aborde la totalidad del período- es doble: por una parte, la homogeneidad territorial de los fondos de estos años, con excepción del período inicial (1768-1774)[4]; por otra parte, la también relativa homogeneidad formal, por lo que se refiere a la tipología de libros de registro e incluso a la continuidad de las manos que los escribieron.[5] Atendiendo al volumen de asentamientos contenidos en los libros de 1771 y 1806, estimamos que la cantidad de escrituras registradas entre 1767 y 1805 se sitúa en torno a las 210.000. El registro completo es probable que supere con creces el medio millón.

No se trata del primer acercamiento que hemos realizado a los fondos de los registros de Hipotecas de la región de Girona. De hecho, antes de plantearnos un vaciado sistemático y masivo, habíamos trabajado en distintos momentos con estos fondos y lo habíamos hecho a través de dos estrategias complementarias. La más antigua (Congost, 1989, 1990) consistió en realizar una selección, por un lado, de algunas tipologías documentales (establecimientos enfitéuticos) y, por otro, de algunos contratos en que las élites eran protagonistas (escrituras de compraventas de fincas con un valor superior a las 1.000 libras, escrituras de contratos matrimoniales con dotes superiores a esta misma cantidad). Estos registros se vaciaron sistemáticamente para todos los años de existencia del registro, lo cual permitió acumular unas 55.000 entradas referidas a escrituras. Se trata de una aproximación que, si bien dio frutos interesantes, fijaba su atención en los grupos sociales dominantes. La segunda estrategia consistió en vaciar exhaustivamente la totalidad de los asientos correspondientes a un periodo corto. Concretamente se realizaron tres cortes temporales de un año de duración cada uno: 1771, 1806 y 1841. La primera tentativa se limitó al vaciado de una sola tipología, los contratos matrimoniales; pero posteriormente se llevó a cabo el vaciado completo de todos los registros y dio lugar a una base con 18.000 entradas. Algunos de los datos expuestos en el apartado anterior proceden de estos cortes. Esta opción permitió ver las posibilidades que ofrecía un tratamiento masivo y el sesgo que implicaba fijarse solamente en las élites. Sin

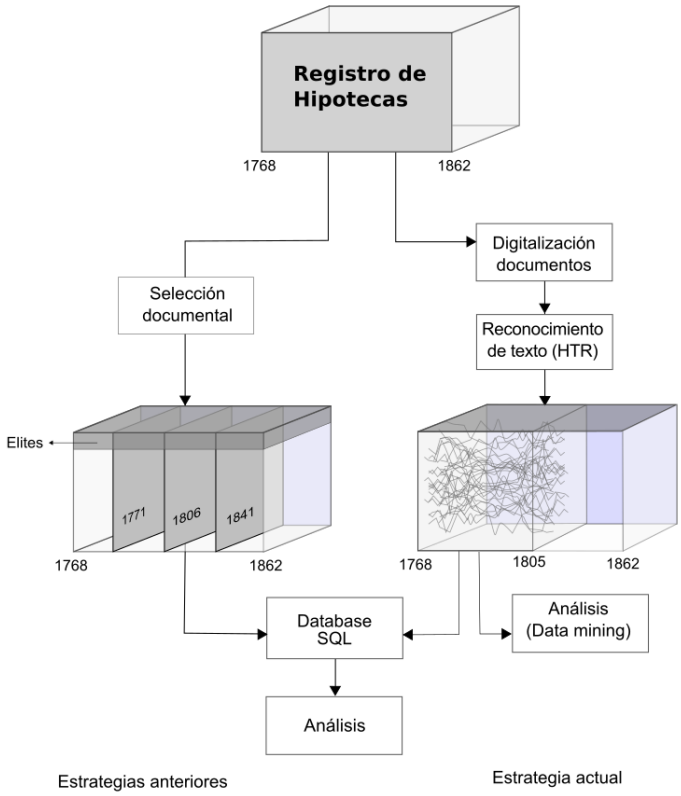
[4] A partir de 1806 disponemos también del fondo del Oficio de Hipotecas de Figueres. Este registro fue creado en 1774, como segregación del Oficio de Girona. Su documentación hasta 1805, sin embargo, está extraviada o desaparecida. La posibilidad de reunir, en fases posteriores, la documentación de ambos oficios a partir de 1806 nos parece un motivo justificado para delimitar el alcance de la primera fase.

[5] Sin haber realizado ningún examen grafológico detallado, ni tampoco haber verificado los cambios en la escribanía encargada del Oficio de Girona, un examen somero permite observar una notable homogeneidad en la caligrafía, mucho mayor que la observable, por ejemplo, en los protocolos notariales.

embargo, al tratarse de períodos muy cortos, se planteaba el problema de cómo observar e interpretar los cambios sociales.

Nuestra propuesta actual de vaciado exhaustivo para un total de 38 años sólo puede llevarse a cabo mediante procedimientos que nos permitan realizarlo con unos costes de tiempo y esfuerzo razonables. Para vaciar las catas temporales de 1771, 1806 y 1841 fue necesario ocupar a seis estudiantes durante dos años a razón de 15 horas semanales. A este ritmo, el vaciado del periodo 1768-1805 significaría 25 años de trabajo. Claramente inviable. Sin embargo, el desarrollo reciente de las tecnologías RTM nos permite ser optimistas y abrigar fundadas esperanzas de poder llevarlo a cabo de forma si no totalmente mecanizada, al menos con una cantidad de trabajo asumible.

Imagen 1. Estrategias de explotación del Registro de Hipotecas



El proceso de tratamiento de la documentación y su conversión en datos manejables tiene cuatro jalones o fases, cada una de las cuales también puede tener sus procesos específicos: (1) fase es la digitalización de los libros del Oficio de Hipotecas, (2) fase de transcripción automática, (3) fase de preprocesamiento analítico y (4) fase de análisis de datos. La última está destinada a desarrollar los objetivos que se han marcado al inicio de este apartado, y sobre la tercera por ahora sólo interesa comentar que, a partir de un corpus textual que, entre 1768 y 1805, estimamos que puede contener unos 14 o 15 millones de palabras, pretendemos dotar al texto plano de una estructura mínima que permita su procesado mediante el lenguaje

de bases de datos (SQL). El resultado de la fase de transcripción (2) es un archivo de texto plano en el cual se han identificado algunos elementos específicos (las regiones del texto, la tipología documental, los antropónimos, los topónimos y los oficios) que han sido etiquetados automáticamente en xml. Ello nos deberá permitir su transformación en una base de datos de tipo relacional, convirtiendo las etiquetas en campos a partir de una serie de reglas estructurales. No queremos, sin embargo, insistir ahora sobre esta cuestión. Nuestro objetivo, en esta comunicación, es centrarnos en la fase de transcripción.

La fuente: los libros del Registro de Hipotecas de Girona y Figueres

El Registro de Hipotecas fue una institución de publicidad registral creada mediante una Real Pragmática de 31 de enero de 1768.[6] Para ser más precisos, la denominación de *Registro de Hipotecas* no surgió hasta su última reforma, en 1845. Originariamente, se impuso el nombre de *Oficio de Hipotecas*, el cual, a su vez, se trocó en 1829 por el de *Contaduría de Hipotecas*. Estos cambios nominales, tras los cuales se ocultan reformas de mayor calado sobre su alcance y organización interna, no impiden que la concibamos como una misma institución cuya continuidad sólo se vio alterada por la aprobación de la ley Hipotecaria de 1861, que en esta ocasión sí implicó una ruptura y su sustitución por el moderno Registro de la Propiedad.

La Real Pragmática de 1768 instituyó el establecimiento de un Oficio en todas las localidades que fueran cabeza de partido, nombrando responsables de los mismos a los escribanos de sus cabildos o ayuntamientos. En los libros de dicho registro debía tomarse razón de los datos básicos de todos los instrumentos notariales *de imposiciones, ventas, y redenciones de censos, ò tributos, ventas de bienes raíces, ò considerados por tales, que constare estar gravados con alguna carga, fianzas, en que se hipotecaren especialmente tales bienes, Escrituras de Mayorazgos, ù Obra pìa, y generalmente todos los que tengan especial, y expressa hipoteca, ò gravamen, con expresion de ellos, ò su liberación, y redención*.[7] Como veremos, la interpretación de este artículo dio lugar variaciones en su contenido que, en el caso catalán, se concretaron en una ampliación a todas las transmisiones de bienes inmuebles por título de compraventa, independientemente de si se hallaban o no gravados o hipotecados.

Los interesados disponían de un plazo relativamente breve para presentar la escritura original en el Oficio, que podía retenerla durante veinticuatro horas. En este lapso, el encargado del registro debía redactar un asiento donde constara la fecha del instrumento, el nombre y vecindad de los otorgantes, la calidad del contrato –indicando el tipo de gravamen-, y una descripción de los bienes afectados, con expresión de su situación, cabida y lindes. Según dicha Pragmática, debía llevarse *un Libro, ò en muchos, registros separados de cada uno de los Pueblos del Distrito* , aunque en la práctica no parece que este criterio fuera de aplicación

[6] *Pragmática Sanción de su Magestad, en fuerza de ley, en la qual se prescribe el establecimiento del Oficio de hipotecas en las Cabezas de Partido al cargo del Escribano de Ayuntamiento para todo el Reyno, y la Instrucción que en ellos se ha de guardar, para la mejor observancia de la Ley 3. tit. 15. lib. 5. de la Recopilación, con lo demás que expressa*, 31 enero 1768. Sobre los antecedentes tanto en el ámbito de la Corona de Castilla como de la Corona de Aragón, ver Serna (1995).
[7] *Pragmática Sanción…*,31 enero 1768, art. XVI.

habitual, ni por lo que conocemos de los registros catalanes (Congost Colomer, 1990; López, 1974; López & Tatjer Mir, 1986), ni por lo referido a otras regiones (Cerdeña Ruiz, 2003). Los libros de registro de ámbito municipal, en el caso de Girona y de otras oficinas catalanas, no aparecieron hasta la 1845 (Canela i Garayoa, 1985; Congost Colomer, 1990), cuando un decreto redefinió el contenido e impuso la municipalización de los libros.

Las oficinas del registro: una geografía cambiante

Una primera dificultad que cabe sortear cuando se plantea un vaciado sistemático de los libros de los Oficios de Hipotecas son los cambios tanto en el número de oficios existentes como en su distribución territorial. Inicialmente, la Real Pragmática de 1768 estableció la creación de una oficina en todos los municipios que fueran Capital o Cabeza de Partido. Se trataba de un criterio genérico que debía ajustarse a las formas y denominaciones de las divisiones territoriales existentes en cada lugar. Cataluña no se organizaba en partidos, sino en corregimientos, y en aquel momento existían siete. La traslación de la Pragmática al corregimiento de Girona dio lugar a dos oficinas iniciales: la de Girona y la de Besalú. La primera era capital corregimental, y la segunda sede de una Alcadía Mayor (Consejo de Estado, 1789; Nomenclator, 1789; Serramontmany, 2016, pp. 27–38, 321). Sin embargo, enseguida surgieron peticiones de diversas localidades solicitando la concesión de un oficio de hipotecas. Sebastià Villalón (2008) ha documentado, para Cataluña, un total de veinte peticiones, de las cuales tres correspondían al corregimiento de Girona. Dos de ellas la Real Audiencia de Cataluña las resolvió favorablemente en 1774, con lo cual se desgajaron del territorio inicial las oficinas de Figueres y de Hostalric (Villalón, 2008). La buena recepción que, en Cataluña, tuvo esta institución registral, la rápida saturación de las primeras oficinas que se crearon y la voluntad superar los obstáculos orográficos y de comunicación impulsaron la multiplicación de los oficios y, por ende, la redefinición territorial de los iniciales.

La historia archivística de estas cuatro oficinas tuvo recorridos distintos. Por una parte, tenemos los libros de las dos oficinas iniciales (Girona y Besalú), que actualmente se hallan depositados en archivos distintos (el Arxiu Històric de Girona, en el primer caso; el Arxiu Comarcal de la Garrotxa, en el segundo). Por otra, tenemos que los libros del Oficio de Hipotecas de Girona entre 1768 hasta 1774 contienen los registros de las escrituras de los territorios que luego corresponderían a los oficios de Figueres y Hostalric. La mezcla de asientos referidos a todas las localidades en un mismo libro general impidió la separar la documentación de los primeros años. Quizás fuera una suerte, ya que la serie de libros de Figueres, que debía empezar en 1774, está perdida hasta el año 1806. A partir de esta fecha y hasta su extinción, el fondo se halla en el Arxiu Històric de Girona. Finalmente, los libros del Oficio de Hostalric también están extraviados, aunque una nueva división de este antes de 1780 dio lugar al Oficio de Calella, cuya documentación se conserva en el Arxiu Fidel Fita de Arenys de Mar (Burgueño, 2004, p. 14)

Con posterioridad, el distrito hipotecario de Girona continuó experimentado divisiones, aunque en este caso, por razones que por ahora desconocemos, no se han traducido en un creciente desgajamiento de los fondos documentales, que se conservan unificados en el Arxiu

Històric de Girona. Las nuevas contadurías aparecieron a raíz o con posterioridad a la reforma de 1829, y fueron las siguientes: Banyoles (1829-1835), La Bisbal (1836), Sant Feliu de Guíxols (1831) y Torroella de Montgrí (1829).[8]

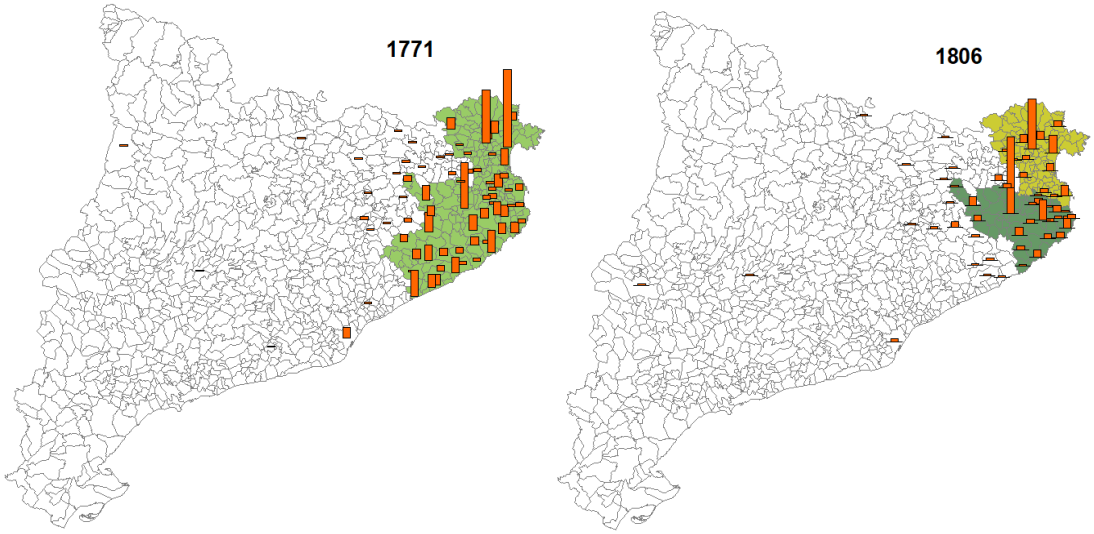
Tabla 1. Volúmes del Registro de Hipotecas conservados en el Arxiu Històric de Girona, por Contaduría y tipología de libros

Oficina	registro general	registro particular	traslados de dominio	arrenda-mientos	otros	complementos	índices	nº volúmenes
Banyoles	2 (1829-1835)	1 (1831-1835)						3
Besalú	6 (1768-1872)		13 (1679-1904)		1 (1841-1842)		4 (1768-1862)	24
Castelló d'Empúries	22 (1829-1862)	6 (1831-1861)	57 (1701-1868)		1 (1851-1854)		19 (1747-1862)	105
Figueres	55 (1806-1862)	14 (1830-1862)	92 (1693-1867)	36 (1815-1865)		5 (1845-1878)	36 (1806-1862)	238
Girona	154 (1768-1862)	18 (1830-1847)	104 (1845-1862)	34 (1819-1866)	28 (1768-1877)		150 (1768-1868)	488
Hostalric	13 (1809-1841)							13
La Bisbal	44 (1828-1862)	6 (1831-1846)	70 (1692-1905)	8 (1842-1862)			8 (1800-1829)	136
Sant Feliu de Guíxols	8 (1828-1857)	10 (1831-1847)	30 (1725-1871)	2 (1847-1862)			1 (1845-1855)	51
Santa Coloma	37 (1800-1870)	1 (1866-1866)	77 (1604-1889)	12 (1840-1886)	14 (1300-1866)	2 (1866-1891)	4 (1857-1870)	147
Torroella de Montgrí	16 (1829-1862)	10 (1829-1862)	48 (1838-1884)	13 (1838-1862)	2 (1806-1861)		4 (1829-1860)	93
nº volúmenes	357	66	491	105	46	7	226	**1298**

Fuente: Arxiu Històric de Girona, fondos registrales. Notas: (1) Las fechas extremas proceden de la descripción archivística de cada volumen, que en algunas ocasiones remite a escrituras anteriores a la creación del Oficio de Hipotecas. (2) Las tipologías agrupan las distintas denominaciones con que se titularon los libros.

La documentación que, en la primera fase del proyecto, tenemos la pretensión de transcribir y analizar alcanza hasta el año 1805 y se circunscribe al territorio correspondiente al Oficio de Girona. Excepto para los primeros seis años, cuando aún estaban incorporadas los futuros oficios de Figueres, Hostalric y Calella, los problemas que pueda generar esta geografía cambiante sobre la homogeneidad de la serie son fácilmente subsanables. Mayores dificultades pueden plantearse cuando el proyecto, en fases posteriores, vaya más allá de 1829.

Imagen 2. Distribución de las escrituras registradas, según la localidad donde se ubicaba la notaría



[8] Entre paréntesis indicamos la fecha del primer volumen de la serie correspondiente a cada Contaduría. En el caso de Banyoles, la serie sólo alcanza hasta 1835.

Otro aspecto relevante, aunque de índole distinta a los cambios territoriales, es la obligación, establecida ya en la Pragmática de 1768, de presentar las escrituras no en los oficios más próximos a la notaría donde se habían escriturado, sino en aquellos donde se hallaran los bienes afectados. Esta norma comportaba una corrección a la libertad en la contratación notarial que daba opacidad a los tratos y, desde la perspectiva del investigador, permite una información más completa a nivel local que la que puede obtenerse de las notarías. La importancia de esta práctica es doble, tal y como ilustra la secuencia de mapas: (a) Permite controlar las escrituras realizadas fuera del distrito hipotecario. Aunque, en la representación cartográfica, su volumen parece menor y con tendencia decreciente, no debe ser menospreciado. En 1771 hubo 65 notarios externos al distrito hipotecario que escrituraron contratos referidos a este territorio. Esta cifra, posteriormente, decayó -como también lo hizo el número de notarios locales- pero continuó siendo importante (28 notarios en 1806 y 46 en 1841).

(b) También nos permite superar las distorsiones en la distribución geográfica de la actividad notarial. Como se observa en la secuencia de mapas, mucho antes de la reorganización del mapa notarial que impuso la Ley del Notariado de 1862, la actividad ya había tendido a concentrarse en un número reducido de localidades y, además, los polos de concentración habían experimentado variaciones significativas (en particular el relevo de Castelló d'Empúries por Figueres). No cabe descartar que la propia geografía del registro de hipotecas hubiera contribuido a ello. Lo que nos importa remarcar, en todo caso, es que cualquier aproximación a través de una notaría o de una muestra de notarías podría verse afectada por estos movimientos, a diferencia de una mirada desde el Oficio de Hipotecas.

Características de la fuente: contenido

La información sobre propiedad inmobiliaria que contiene el registro de hipotecas es parcial, dado que, en principio, estaba limitada a aquellos contratos que contuvieran cargas e hipotecas sobre propiedades. Ello imprime un primer sesgo relevante a cualquier posibilidad de explotación analítica de dichos fondos. Sin embargo, es preciso destacar que los criterios que definieron el contenido del registro no parecen haber sido ser homogéneos a lo largo del territorio español. Numerosos autores destacan, en esta línea, las peculiaridades de las oficinas catalanas. Marina López indicó que *las reformas que, en 1774, la Real Audiencia introdujo en la institución para adaptarla a la constitución jurídica del Principado, la hicieron sustancialmente distinta de sus homónimas en el resto del país* (López, 1974). También Margarita Serna (1995, pp. 283–286) sostiene que, en Cataluña, al no haber existido durante la edad moderna una legislación propia en esta materia y haber predominado un régimen de clandestinidad inmobiliaria, cuando se publicó la Pragmática de 1768 las instituciones y los colegios notariales no sólo no se opusieron a la creación de los Oficios de Hipotecas, sino que lo impulsaron tanto en lo que se refiere la ampliación de su objeto como del número de Oficios. Las resoluciones del Consejo de Castilla en 1769, respondiendo a las dudas trasladadas por la Real Audiencia de Cataluña, y el edicto emitido en 11 de julio de 1774 por esta segunda

institución resolvieron algunas dudas habían planteado los notarios referidas a si debían registrarse determinados tipos contractuales, en particular aquellos que contenían una hipoteca de carácter general sobre todos los bienes de alguna de las partes contratantes, los que contenían obligaciones temporales que se extinguían con un recibo simple (debitorios, arrendamientos,…) y los testamentos que no contenían vínculo o gravamen (*hereu gravat*). Las resoluciones implicaron la inclusión en el registro de un amplio número de tipologías documentales. Se dictaminó que debían inscribirse todas las escrituras que contuvieran hipotecas generales, todos los documentos donde existiera algún fideicomiso, se abrió la puerta a la inscripción de debitorios, censos consignativos, establecimientos enfitéuticos, capítulos matrimoniales, arrendamientos,… y también se incluyeron todas las ventas de bienes raíces, aunque no estuvieran gravadas con hipotecas (Villalón, 2008). El resultado final fue un registro que, si bien sólo contenía una parte de la contratación notarial, se trataba de una parte sustancial.

Tabla 2. Tipologías documentales básicas en los Oficios de Hipotecas de la región de Girona, 1771, 1806 y 1841

	1771	**1806**	**1841**
Operaciones crédito	1.797	1.093	1.499
Compraventa	978	777	1.286
Establecimientos	422	243	806
àpocas, cartas de pago	155	506	630
Arrendamientos	405	781	155
donaciones, cessiones	159	146	280
definiciones	21	76	43
modificaciones de precio	27	57	20
reducciones	65	6	3
poderes	3		
heredamientos, capítulos,...	53	139	162
Inventarios	54	25	15
cabrevaciones, confessiones	4	1	4
otros	484	424	552
	4.627	**4.274**	**5.455**

Nota: Utilizamos como marco de referencia la "región de Girona", que incluye todos los Oficios que se segregaron del Oficio de Girona con posterioridad a 1768 cuyos fondos se encuentran depositados en el Arxiu Històric de Girona, esto és: Figueres, Sant Feliu de Guíxols, Torroella de Montgrí, Banyoles y La Bisbal. Se excluyen los Oficios de Hostalric y Calella.

La Tabla 2, referida a los contenidos del registro, nos permite visualizar la diversidad de operaciones que estuvieron sujetas a registro en tres momentos concretos. Como era de esperar, la primera posición la ocupan las operaciones de crédito: creación, transmisión (*encarregament*) y extinción (*lluïció*) de censos consignativos, ventas a carta de gracia, debitorios, insolutumdaciones,… El segundo capítulo más abundante lo constituyen operaciones mercantiles que implicaban transmisión del dominio, básicamente compraventas y las distintas modalidades de establecimiento enfitéutico. El efecto de las resoluciones adoptadas entre 1769 y 1774 se hace patente. Es interesante remarcar, también, la abundancia de escrituras de pago (ápocas, cartas de pago), hecho poco habitual en otras oficinas (Díaz Capallera, 2012), y también un notable volumen de escrituras de arrendamiento. Por el contrario, hay tipologías bastante frecuentes en los protocolos notariales que o no aparecen o solo lo hacen de forma excepcional. Los contratos de poderes o de nombramiento

de administradores y apoderados, por ejemplo, son poco habituales en el registro. Igualmente no es habitual encontrar la documentación judicial (diligencias, requerimientos,….) presente en muchos protocolos, excepto en el caso de las concordias. En la Oficina de Girona de 1771 se registraron 94 escrituras de concordia, una cifra nada despreciable.

Si se desciende a comparar a escala catalana el contenido de los libros de registro de distintas oficinas, pueden hallarse diferencias significativas entre ellas. En parte son explicables por la práctica notarial dominante en cada ámbito territorial y por el tipo de negocios que acababan contratándose en las mesas de los notarios, lo cual a su vez es reflejo de la actividad económica de cada zona; pero probablemente también influyeron otros factores como el grado de confianza del colectivo notarial o de los contratantes en el registro de hipotecas como instrumento para asegurar sus derechos jurídicos (Villalón, 2008). Disponemos de algunos datos que ilustran esta diversidad de contenidos. Una comparación entre el abanico de tipologías documentales contenidas en dos oficios gerundenses (Girona y Figueres) y dos oficios tarraconenses (Montblanc y Tarragona) realizado por Eduard Díaz (2012), pone de relieve que los registros gerundenses incluían un mayor número de tipos documentales, de lo que puede deducirse una concepción más amplia de lo que debían registrar (Tabla 3). También se observa una reducción de este abanico entre las dos fechas analizadas, que interpretamos más en términos de estandarización de la nomenclatura que de modificación de los criterios de registro. En cualquier caso, las diferencias entre ambas regiones parecen mantenerse. También es indicador de una probable existencia de prácticas diferenciadas entre los oficios catalanes el distinto peso que determinadas escrituras o tipos contractuales tenían en cada uno. En este sentido, es relevante el elevado peso que tenían las operaciones de traslación de dominio (compraventas, establecimientos,…) en comparación con las tipologías crediticias o los contratos matrimoniales (Tabla 4).

Tabla 3. Tipologías documentales en distintos Oficios de Hipotecas, 1773 y 1807, en número de tipos distintos

	1773	1807
Girona	154 (5756)	97 (3095)
Figueres	-	73 (2158)
Montblanc	44 (1318)	45 (909)
Tarragona*	76 (3659)	66 (2088)

Fuente: Díaz Capallera, 2012, pp. 13–18. Entre paréntesis se indica el número de documentos registrados durante el año correspondiente. El Oficio de Tarragona (*), en 1773, incluía el territorio que posteriormente (y antes de 1807) dio lugar al Oficio de Reus.

Tabla 4. Peso de las principales categorías de operaciones inscritas en distintos Oficios de Hipotecas, 1773 y 1807

	1773			1807		
	Girona	Montblanc	Tarragona	Girona	Montblanc	Tarragona
Traslaciones de dominio	25,8%	52,7%	22,2%	27,8%	72,0%	41,3%
Operaciones de crédito	27,7%	6,7%	18,1%	14,9%	8,3%	25,1%
Capítulos matrimoniales	9,6%	18,1%	17,2%	15,9%	7,3%	8,9%

Fuente: Díaz Capallera, 2012, pp. 29–30

Algunos autores han realizado distintos ejercicios de valoración del grado de clandestinidad inmobiliaria que afectó a los Oficios de Hipotecas. Marina López y Mercè Tatjer comprobaron, en el Oficio de Barcelona, el registro de escrituras notariales referidas a sociedades mercantiles, llegando a la conclusión que la clandestinidad era baja y la fiabilidad del registro alta. Sebastià Villalón, por su parte, ha contrastado los contratos o capítulos matrimoniales presentes en 1775 en nueve oficios catalanes con la tasa de nupcialidad, llegando a una conclusión semejante.

Tabla 5. Días transcurridos entre la escritura notarial y su registro en los Oficios de Hipotecas de la región de Girona, 1771, 1806 y 1841

	% sobre las actas de cada año		
Días transcurridos	1771	1806	1841
< 31	77,8%	88,7%	90,9%
31-60	7,5%	3,8%	2,3%
61-91	0,4%	0,2%	0,4%
91-182	0,7%	0,2%	0,4%
183-364	0,6%	0,2%	0,4%
365-728	0,5%	0,3%	0,1%
729-1096	0,4%	0,1%	0,1%
>3 años	12,1%	6,4%	5,3%
	100%	100%	100%
Nº actas	5.258	4.566	5.758

Las distintas normativas establecieron plazos relativamente breves para cumplir con la obligación de presentar las escrituras en los Oficios de Hipotecas correspondientes. En 1771 el 78% de las escrituras se registraron antes de superar el mes respecto a la fecha de cierre del acta notarial; en 1806 ya era el 89% y este porcentaje se incrementó hasta el 91% en 1846 (Tabla 5).

Tabla 6. Fecha de escritura de las actas notariales registradas en los Oficios de Hipotecas de la región de Girona, 1771, 1806 y 1841

periodo	1771	1806	1841
s. XIII	12	1	0
s. XIV	9	6	0
s. XV	8	3	1
s. XVI	27	24	3
s. XVII	177	77	24
1701-1750	258	106	73
1751-1767	140	27	47
1768/71- 1802	4.627	49	55
1803/06- 1837	0	4.273	101
1838-1841	0	0	5.454
Nº actas	5.258	4.566	5.758
% actas antiguas	12,0%	6,4%	5,3%

En el caso de escrituras otorgadas con anterioridad a la Pragmática, también se estableció la obligatoriedad de su registro, siempre que las cargas aún estuvieran vigentes. En 26 de febrero 1774 el Consejo de Castilla emitió una circular en que recordaba la obligación de inscribir

aquellas que contuviesen cargas o censos y se dejó primer un plazo de dos meses para cumplirlo; luego, en 1 julio de 1774, el plazo se alargó hasta un año (Cerdeña Ruiz, 2003). Es probable que ello diera lugar a un aumento puntual del registro de escrituras antiguas, aunque aún no podemos verificarlo. Lo que sí se observa es un degoteo relativamente constante de inscripciones relativas a actas antiguas o, incluso, muy antiguas. En algún caso las escrituras databan de la época medieval, aunque la mayoría correspondían al periodo más cercano. Los decretos ordenando la necesaria inscripción de todas las escrituras con cargas vigentes no parece que hubieran dado el fruto esperado y, en cualquiera caso, parece que, a pesar de las amenazas, el registro permaneció siempre abierto a la inscripción de escrituras antiguas, e incluso de las que se habían realizado con posterioridad a 1768. Naturalmente los primeros años de constitución del registro, el volumen de actos antiguos fue mayor (del orden del 12% de los asientos); sin embargo, el flujo que se observa con posterioridad, del orden de un 5-6% anual, no parece una cifra despreciable. El diferencial entre fecha notarial y fecha registral es una característica de la fuente que debe tenerse muy en cuenta en el análisis de los datos, para evitar mezclar documentos que, aunque registrados al mismo tiempo, se refieren a hechos potencialmente distantes en el tiempo.

Con todas las prevenciones sobre la continuidad geográfica, la correspondencia cronológica y la selección tipológica que hemos expuesto, creemos que los fondos del Registro de Hipotecas, al menos los correspondientes a la región de Girona constituyen una serie documentalmente rica, suficientemente continua y territorialmente bastante amplia, y que, por ello, tiene un enorme potencial para el análisis de los cambios sociales. Veamos, a continuación, hasta qué punto los trabajos preliminares de transcripción asistida realizados hasta el momento hacen posible esperar que el proyecto sea viable.

El reconocimiento de texto manuscrito aplicado a los libros del Registro de Hipotecas de Girona

Las pruebas preliminares y de entrenamiento del sistema de reconocimiento de texto manuscrito llevadas a cabo hasta el momento se han realizado con una parte del segundo libro del *Registre general d'Escriptures* del Oficio de Hipotecas de Girona, correspondiente al año 1769.[9] Se trata de un volumen bien conservado, que contiene 1179 folios con una gran uniformidad en el trazo manuscrito, y que fue digitalizado por el Arxiu Històric de Girona en el marco del Pla Bruniquer. Como se expone a continuación, el producto resultante del proceso de transcripción es un archivo que, además del texto sin formato, también contiene algunas etiquetas xml de especial interés para su posterior manipulación, ya que permiten identificar los antropónimos, los topónimos, las fechas y las categorías sociolaborales.

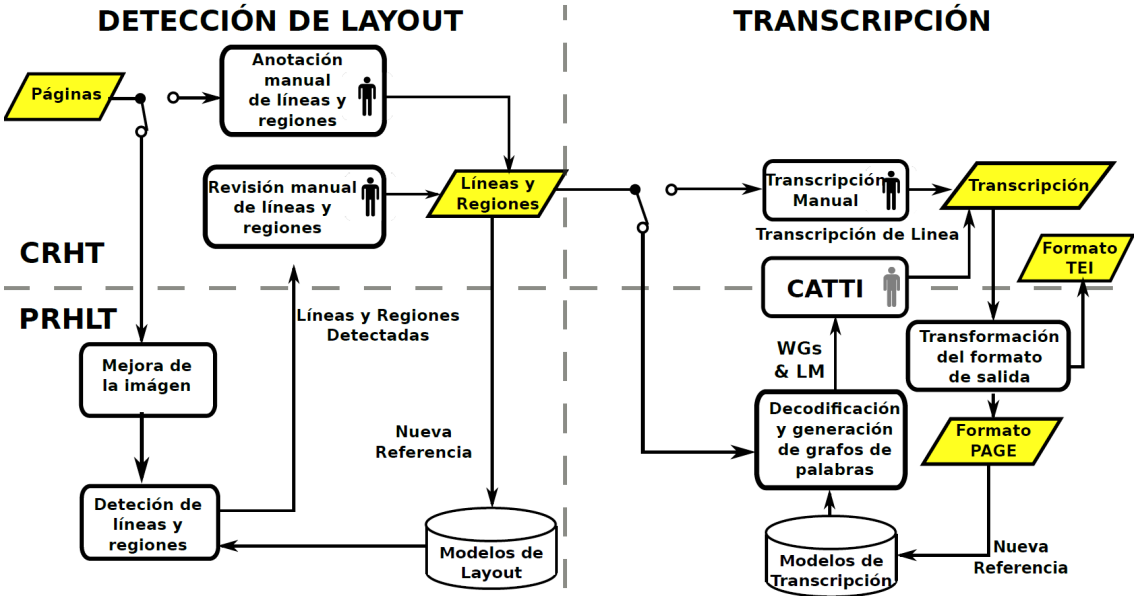
[9] Arxiu Històric de Girona, Comptaduria d'Hipoteques de Girona, 2. Puede accederse a la versión digital a través de http://arxiusenlinia.cultura.gencat.cat.

Flujo de trabajo definido para la transcripción de texto

Debido a la larga extensión de la colección y la necesidad de obtener una transcripción perfecta, el proceso se ha llevado a cabo de manera interactiva en lotes de 50 páginas cada uno. En cada bloque se sigue el flujo de trabajo descrito en la

Imagen 3.

Imagen 3 Flujo de trabajo utilizado para la transcripción asistida



Cada conmutador indica un punto en el flujo donde es posible pasar de un proceso manual a uno automatizado, utilizado principalmente al inicio del proyecto cuando no hay suficientes datos para entrenar los modelos

El proceso de transcripción se inició con un lote inicial de 50 páginas (el cual llamaremos b001), anotado manualmente utilizando el software Transkribus[10], tanto para el *layout* como para la transcripción. Este lote fue utilizado para entrenar los modelos de RTM y así obtener la transcripción del siguiente bloque (b002).

Seguidamente el usuario corrige cualquier error presente en la transcripción por medio del software de transcripción asistida CATTI[11], tal como se muestra en la Imagen 4. La transcripción obtenida en este paso se utiliza para re-entrenar los modelos de RTM y de esta forma mejorar las prestaciones de los mismos.

El proceso se repitió para los próximos 5 lotes (b002-b006). A partir del bloque b007 se introdujo una herramienta de análisis automático del documento (*layout*), la cual permite extraer de manera semi-automática la segmentación de las líneas de texto y posteriormente

[10] https://transkribus.eu/Transkribus/
[11] http://transcriptorium.eu/demots/htr/index.php

las diferentes regiones de texto presentes en el documento, en la Tabla 7 se muestran las diferentes regiones definidas y su identificador.

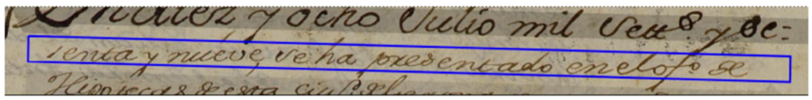
Tabla 7 Etiquetado de las regiones de texto.

Identificador	Descripción
pag	Número de página
tip	Tipología Documental
par	Párrafo que inicia junto a su tipología documental
pac	Párrafo que inicia en una página anterior
not	Nota al margen del texto
nop	Nota al margen, agregada a posteriori

De esta manera se facilita el proceso de anotado del *layout* para el usuario y por ende se reduce el tiempo invertido en el mismo. Las imágenes son anotadas por el sistema automático y posteriormente editadas por el usuario para corregir cualquier error.

Al igual que en el caso de las transcripciones, las imágenes de un lote revisadas por el usuario son utilizadas para re-entrenar los modelos utilizados en la anotación automática del siguiente lote.

Imagen 4 Ejemplo de uso de CATTI, con dos interacciones del usuario para obtener la transcripción



→- $tip:Venta y nueve se ha presentado en el of.^{o}.oficio de
→, $-senta y nueve se ha presentado en el of.^{o}.oficio de
$-senta y nueve, se ha presentado en el of.^{o}.oficio de

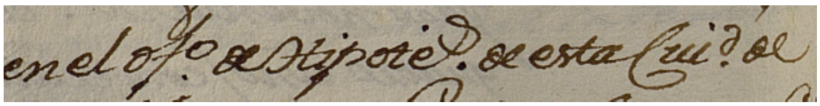
Nota: El usuario escribe un guion después de la primera aparición de "$"; luego el usuario escribe una coma después de la palabra "nueve".

Criterio de transcripción: antes de iniciar la tarea de transcripción, algún criterio debe ser establecido respecto a la forma de realizarlo. El tipo más común de transcripción para documentos históricos manuscritos utilizado por los paleógrafos es la llamada transcripción diplomática o paleográfica, la cual trata de transcribir tipográficamente lo más exacto posible todas las características importantes del manuscrito original, incluido ortografía, signos de puntuación, abreviaturas, tachaduras, inserciones, y cualquier otra alteración (Bosch et al., 2014). En éste trabajo hemos adoptado las reglas de la transcripción diplomática, con algunos cambios menores para adaptarse a la tecnología de RTM utilizada.

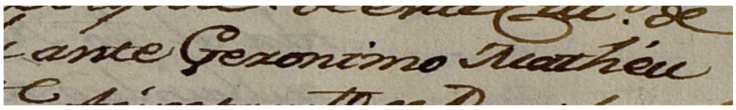
Adicionalmente, las transcripciones fueron etiquetadas con información enriquecida/complementaria (expansión de las abreviaturas, marcas de división silábica, etc.), inicialmente las etiquetas fueron agregadas manualmente, y a partir del lote b003 son agregadas automáticamente en el proceso de RTM. Para facilitar el proceso de etiquetado se

definió un conjunto adecuado de etiquetas, utilizando el carácter especial "$". Dichas etiquetas pueden ser fácilmente convertidas a formato TEI u otro formato. La lista completa de etiquetas y su significado se muestra en la Tabla 8 y un ejemplo de transcripción utilizando dichas etiquetas en la Imagen 5.

Imagen 5 Ejemplo de transcripción con etiquetas.



en el of.**^{o}.oficio** de Hipote.**^{s}.Hipotecas** de esta Ciu.**^{d}.Ciudad** de



ante **$ant:**Geronimo **$ant:**Mateu

Tabla 8 Etiquetas permitidas en el proceso de transcripción.

Descripción	Etiqueta(ejemplo)
División silábica	Pre**$-**
Continuación de división silábica	**$-**sentado
Sobre-índice	Hipote.**$^**s
Abreviatura	Nro**$.Nuestro**
Tachadura/ilegible	**$#**
Antropónimo	**$ant:**Mateu
Nota post-edición	**$nop:**palabra
Oficio	**$ofi:**Labrador
Número de página	**$pag:**555
Tipología Documental	**$tip:**Venta
Topónimo	**$top:**Gerona

Experimentación

Más de 590 documentos (12 lotes) de ésta colección fueron procesados con el sistema semi-automático descrito anteriormente, y registrados en formato PAGE-XML y TEI. Los documentos procesados hasta el momento se componen de más de 23700 líneas de texto, cubriendo más de 50 temas diferentes (tipologías documentales) y un vocabulario de más de 2400 palabras. Los resultados por cada lote se presentan en la siguiente sección.

Resultados

Los resultados desde el punto de vista cualitativo se reportan en término de las métricas estándar de facto para RTM:

- Tasa de error de caracteres (en inglés Character Error Rate –CER–): es una medida común de la calidad de la transcripción. Se define como el número de errores presentes en la transcripción automática (en la forma de inserciones, borrados y

substituciones a nivel de carácter) dividido por el número de caracteres presentes en la transcripción correcta.

- Tasa de error de palabras (en inglés Word Error Rate –WER–): es una medida similar al CER, pero se calcula a nivel de palabras.
- Relación de corrección de palabra (en inglés Word Stroke Ratio –WSR–): es un estimador del esfuerzo necesario, por parte de un transcriptor humano, para producir la transcripción correcta. WSR se define como el número de interacciones, con el sistema, que el usuario debe realizar para obtener la transcripción de referencia de la imagen de texto considerada, dividida por el número de palabras presentes en la transcripción de referencia.
- Estimación de la reducción del esfuerzo (en inglés Estimated Effort Reduction –EFR–): se define como la diferencia relativa entre el WER y el WSR. Nos proporciona una estimación de la reducción en el esfuerzo del usuario, necesario para obtener la transcripción correcta, al utilizar el sistema de transcripción asistida respecto a utilizar un sistema de RTM convencional seguido de un proceso de post-edición.

En la Imagen 6 se reportan los resultados obtenidos a nivel del CER de transcripción de cada uno de los 11 lotes procesado hasta el momento, además se reporta el WER con (WER+$) y sin etiquetas.

Imagen 6 Resultados obtenidos por cada lote procesado.

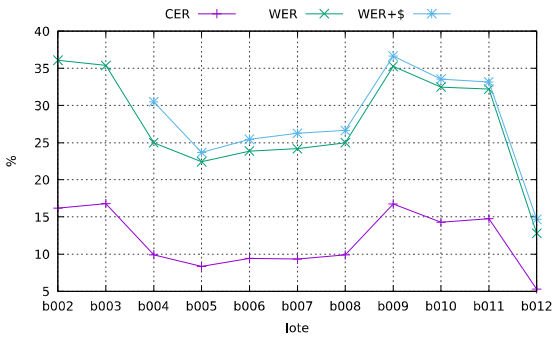
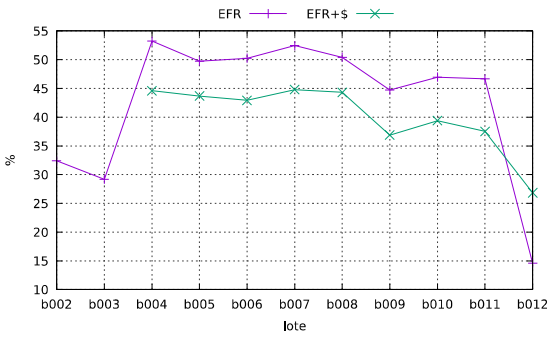


Imagen 7 Estimación de la reducción de la reducción del esfuerzo (EFR) para cada lote procesado.



Nota: El símbolo "+$" indica que la métrica fue calculada con las etiquetas

Inicialmente podemos constatar una disminución paulatina en el número de errores para los primeros lotes, hasta el lote b008, los resultados se mantienen relativamente estables. A partir del lote b009 sospechamos se da cambio en los documentos que incrementan la dificultad de ser transcritos, pero no hemos sido capaces hasta el momento de identificar específicamente cual es.

Para el lote b012 se introdujo un nuevo modelo óptico basado en redes neuronales profundas, que supone una mejora, pese al aumento en la dificultad experimentado en los lotes previos, superior al 34% respecto al mejor resultado obtenido previamente (b005 respecto a b012).

En la Imagen 7 podemos observar como aun cuando la transcripción automática produce errores, el sistema de transcripción asistida (CATTI) logra una reducción de hasta el 53% en el esfuerzo necesario para corregir dichos errores. Cabe destacar que, si el sistema produce resultados con muy buena calidad, como es el caso del lote b012, la reducción en el esfuerzo es mínima debido a que los errores a corregir son muy pocos.

Conclusión

El desarrollo de las tecnologías de reconocimiento de texto manuscrito en los últimos años ha abierto expectativas enormemente atractivas para los historiadores. Aunque los resultados obtenidos contengan cierto margen de error, estos sistemas pueden alcanzar unas cotas de calidad muy notables, especialmente cuando se trabaja con series o colecciones documentales relativamente homogéneas en lo que se refiere a las tipologías y a la caligrafía. Las pruebas realizadas hasta el momento con los libros del Oficio de Hipotecas de Girona permiten ser razonablemente optimistas respecto a la posibilidad de realizar transcripciones masivas con un margen de error aceptable y unos costes de corrección también asumibles. Desde la implementación de un modelo basado en redes neuronales, la capacidad de reconocimiento del sistema desarrollado por el equipo *Pattern Recognition and Human Language Technology* de la Universidad Politécnica de Valencia ha experimentado una mejora considerable y, en el estadio actual, el principal reto lo constituye la reducción de errores en la detección automática de las líneas de texto (layout).

La posibilidad de reconocer y etiquetar algunos elementos (topónimos, antropónimos, tipologías documentales,…) en el momento de la transcripción aporta información adicional que, en fases posteriores, facilitará la manipulación y el análisis de los datos. Por el momento, el resultado obtenido es un archivo de texto plano que contiene el texto transcrito con las etiquetas xml especificadas y que permite búsquedas mediante herramientas convencionales y técnicas de minería de datos.

Bibliografía

Bluche, T. (2015). *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*. Université Paris Sud - Paris XI.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H., & Vidal, E. (2017). Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (p. (01) 311–316).

Bosch, V., Bordes-Cabrera, I., Muñoz, P. C., Hernández-Tornero, C., Leiva, L. A., Pastor, M., … Vidal, E. (2014). Computer-assisted transcription of a historical botanical specimen book: Organization and process overview. In *DATeCH'14. Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 125–130). New York: ACM.

Burgueño, J. (2004). La Tordera. De límit provincial a eix vertebrador d'una comarca. *Matinals. Quaderns D'història Local*, *2*, 11–28.

Canela i Garayoa, M. (1985). Inventari de l'Àntic Registre d'Hipoteques de Cervera. *Miscel·lània Cerverina*, *3*.

Cerdeña Ruiz, R. (2003). La Contaduría de Hipotecas de Fuerteventura : referencias históricas e inventario de su fondo documental. *Tebeto : Anuario Del Archivo Histórico Insular de Fuerteventura*, *16*, 415–485.

Congost, R. (1989). *Els Propietaris i els altres : anàlisi d'unes relacions d'explotació. La regió de Girona, 1768-1862*. Universitat Autònoma de Barcelona, Bellaterra.

Congost, R. (1990). *Els Propietaris i els altres. La regió de Girona, 1768-1862*. Vic: Eumo Editorial.

Congost, R. (2014). Més enllà de les etiquetes. Reflexions sobre l'anàlisi dels grups socials humils. La regió de Girona (1770–1850).

Recerques. Història, Economia, Cultura, *68*, 165–191.

Congost, R., Ros, R., & Saguer, E. (2016). Beyond life cycle and inheritance strategies: The rise of a middling social group in an ancien régime society (catalonia, eighteenth century). *Journal of Social History*, *49*(3). https://doi.org/10.1093/jsh/shv056

Congost Colomer, R. (1990). Una font poc utilitzada: el registre d'hipoteques. *Estudis D'història Agrària*, *8*, 201–234.

Consejo de Estado. (1789). *España dividida en provincias è intendencias : y subdividida en partidos, corregimientos ... con un nomenclator ... de todos los pueblos del reyno, que compone la segunda parte : tomo I*. (en la imprenta real, Ed.). [Madrid] : en la imprenta real.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B*, *39*, 1–38.

Díaz Capallera, E. (2012). *El registre d'hipoteques: una font per a la història social. Estudi comparatiu de les regions de Girona i Tarragona*. Universitat de Girona.

Fiel, S., Grüning, T., Gatos, B., Diem, M., & Kleber, F. (2017). cbad: Icdar 2017 competition on baseline detection. In *Proceedings of the International Conference on Document Analysis and Recognition*.

Fischer, A., Keller, A., Frinken, V., & Bunke, H. (2012). Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, *33*(7), 934–942.

Frinken, V., Fischer, A., Manmatha, R., & Bunke, H. (2012). A Novel Word Spotting Method Based on Recurrent Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(2), 211 –224.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. Cambridge / London: MIT Press.

Kim, G., Govindaraju, V., & Srihari, S. (1999). An architecture for handwritten text recognition systems. *International Journal on Document Analysis and Recognition*, *2*(1), 37–44.

López, M. (1974). Una fuente para la historia de Barcelona: el Registro de Hipotecas. *Estudios Históricos Y Documentos de Los Archivos de Protocolos*, *4*, 345–363.

López, M., & Tatjer Mir, M. (1986). *Inventari dels oficis i comptadories d'hipoteques de Catalunya. Vol. 1*. Barcelona: Dept. de Cultura.

Makhoul, J., Schwartz, R., Lapre, C., & Bazzi, I. (1998). A script-independent methodology for optical character recognition. *Pattern Recognition*, *31*, 1285–1294.

Nomenclator. (1789). *Nomenclator o Diccionario de las ciudades, villas, lugares, aldeas, granjas, cotos redondos, cortijos... con expresión de la provincia, partido y término a que pertenecen...* Madrid: Imprenta Real.

Pastor i Gadea, M. (2007). *Aportaciones al reconocimiento automático de texto manuscrito*. Universitat Politècnica de València.

Plamondon, R., & Srihari, S. N. (2000). On-line and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 63–84.

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., … Lladós, J. (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, *46*(6), 1658–1669. https://doi.org/https://doi.org/10.1016/j.patcog.2012.11.024

Romero, V., Toselli, A. H., & Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*. Singapur: World Scientific Publishing Company.

Sánchez, J. A., Toselli, A. H., Romero, V., & Vidal, E. (2015). ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1166–1170). https://doi.org/10.1109/ICDAR.2015.7333944

Serna Vallejo, M. (1995). *La publicidad inmobiliaria en el derecho hipotecario histórico español*. Universidad de Cantabria.

Serramontmany, A. (2016, February). *Nivells de vida, dinàmiques socials i canvi històric. L'àrea de Besalú, 1750-1850*. Universitat de Girona.

Steinherz, T., Rivlin, E., & Intrator, N. (1999). Off-line cursive script word recognition-a survey. *International Journal on Document Analysis and Recognition*, *2*, 90–110.

Toselli, A. H., Leiva, L. A., Bordes-Cabrera, I., Hernández-Tornero, C., Bosch, V., & Vidal, E. (2017). Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription. *Digital Scholarship in the Humanities*, fqw064. https://doi.org/https://doi.org/10.1093/llc/fqw064

Toselli, A. H., Romero, V., Pastor i Gadea, M., & Vidal, E. (2010). Multimodal interactive transcription of text images. *Pattern Recognition*, *43*(5), 1814–1825.

Toselli, A. H., Vidal, E., & Casacuberta, F. (2011). *Multimodal Interactive Pattern Recognition and Applications*. Springer.

Villalón, S. (2008). Els problemas de la informació en una societat d'Antic Règim. Els notaris catalans davant la creació del Registre d'Hipoteques. In *Dels capbreus al registre de la propietat. Drets, títols i usos socials de la informació a Catalunya (segles XIV-XX)* (pp. 241–274). Girona: Documenta Universitaria.