Inteligencia artificial e historia. El mundo privado de los pobladores del Imperio español

Antonio García-Abásolo
Antonio Calvo Cuenca
Universidad de Córdoba
hi1gaaba@uco.es

El trabajo parte de algunas consideraciones generales sobre las posibilidades de información que el historiador encuentra en los autos de bienes de difuntos y destaca la atención que la historiografía americanista está dedicando en los últimos años a los testamentos como base principal de estudios variados y singulares. En un segundo momento, estas reflexiones se relacionan con la novedad tecnológica que supone para el historiador valerse de técnicas de inteligencia artificial para hacer más eficiente y ágil su trabajo. Las dos partes se unen aquí porque nuestro objetivo es aplicar esta tecnología al estudio de testamentos de pobladores de las Indias españolas en los siglos XVI, XVII y XVIII.

Palabras claves: Inteligencia artificial, bienes de difuntos, testamentos.

Nuevas técnicas para los mismos temas

Por medio de este artículo queremos mostrar las posibilidades de la aplicación de técnicas de inteligencia artificial[1] al estudio de la documenta-

[1] Se denomina inteligencia artificial a la rama de las ciencias de la computación dedicada al desarrollo de agentes racionales artificiales, entendiendo como un agente cualquier cosa capaz de percibir su entorno (recibir entradas), procesar tales percepciones y actuar en su entorno (proporcionar salidas). Debe entenderse la racionalidad como una capacidad humana que permite pensar, evaluar y actuar conforme a ciertos principios de optimización y consistencia. Esta disciplina aborda una gran diversidad de problemas, entre los que se encuentran los de percepción y extracción de información, usando diferentes sistemas sensoriales como la visión, el tacto, el oído, etc. Para resolver estos problemas se proponen tareas genéricas (clasificación, reconocimiento,

ción histórica. En esta ocasión, y dentro de un proyecto I+D titulado *Por la muerte a la vida*,[2] nos hemos centrado en el estudio de testamentos de pobladores de las Indias españolas Occidentales y Orientales de los siglos XVI al XVIII. Esto supone ensayar técnicas nuevas en una de las líneas de investigación habitual del Grupo AAF,[3] dedicada al estudio de la emigración a las Indias, y el mundo privado de los pobladores de América y las Filipinas en la época colonial.

Por otra parte, éste es uno de los aspectos destacados en los objetivos del Proyecto de Excelencia *Andalucía y América Latina. Intercambios y transferencias culturales*,[4] en el que estamos participando. Esta sintonía nos ha llevado a organizar actividades en las que se han tratado cuestiones relacionadas de diversa manera con nuestros objetivos. Una de ellas ha sido el seminario titulado *La música de las catedrales andaluzas y su proyección en América*,[5] en el cual el trabajo de historiadores americanistas con musicólogos e historiadores de la música ha puesto de manifiesto la existencia de objetivos comunes en nuestros respectivos intereses de investigación, entre ellos los relacionados con el aporte humano y cultural que los pobladores de la América española y de las islas Filipinas transportaron a través del Atlántico y del Pacífico en los dos sentidos, tanto a la ida como a la vuelta. Una demostración de esta realidad y uno de los frutos de la actividad desarrollada en

identificación, diagnóstico, etc.), con un planteamiento general en diversos dominios (automóvil, medicina, aeronáutica, ciencias sociales). En particular interesa el problema de la extracción de información sobre diferentes fuentes (imágenes médicas, imágenes obtenidas por robots móviles, textos en lenguaje natural). Queremos indicar con todo esto que, aunque los problemas que se abordan pueden ser distintos y en diferentes ámbitos, las técnicas y herramientas de las que se vale esta rama de la computación son las mismas. Nuestro estudio se ha centrado en un problema de extracción automática de información desde textos escritos en lenguajes natural en el terreno de la historia social, pero las técnicas utilizadas no pueden quedar fuera del marco general que marca la inteligencia artificial y, dentro de ella, el área de procesamiento automático del lenguaje natural.

2 Proyecto I+D del Plan Nacional (Ministerio de Educación y Ciencia, Dirección General de Investigación), código HUM2007-64796.

3 Grupo Andalucía-América-Filipinas, parte del Plan Andaluz de Investigación (PAIDI, HUM187).

4 Proyecto de Excelencia de la Junta de Andalucía, Consejería de Innovación, Ciencia y Empresa, código HUM 03215.

5 Celebrado en la Escuela de Estudios Hispano-Americanos (CSIC), Sevilla, febrero de 2009.

estos proyectos es el libro *La música de las catedrales andaluzas y su proyección en América*.[6]

Más centrado en el tema que nos ocupa aquí, hemos desarrollado otro seminario sobre *Aplicación de técnicas de inteligencia artificial a la documentación histórica*,[7] en el que expertos en computación han mostrado que las aplicaciones de utilidad práctica en la actividad empresarial y administrativa, como la clasificación de productos para control de calidad o la detección de correos electrónicos no deseados, pueden ser muy valiosas si se utilizan adecuadamente en el trabajo del historiador. Hemos organizado este seminario para considerar estas posibilidades y para mostrar los objetivos y los primeros resultados del trabajo sobre testamentos de los pobladores españoles de Indias, realizado con la aplicación de estas nuevas tecnologías.

Fuentes y objetivos

El proyecto tiene dos objetivos complementarios. El primero ha sido formar una base de datos de testamentos de pobladores de América y las islas Filipinas y, en algunos casos, de sus parientes en España, correspondientes a los siglos XVI, XVII y XVIII. El segundo es el análisis informático de esos testamentos, teniendo en cuenta que se trata de un tipo de documento de partes bien definidas, de manera que esperamos que nuestro diseño nos proporcione un sistema de información construido a partir de la ingeniería de ontologías, y un procesamiento informático para el análisis histórico y para el estudio del patrimonio documental.

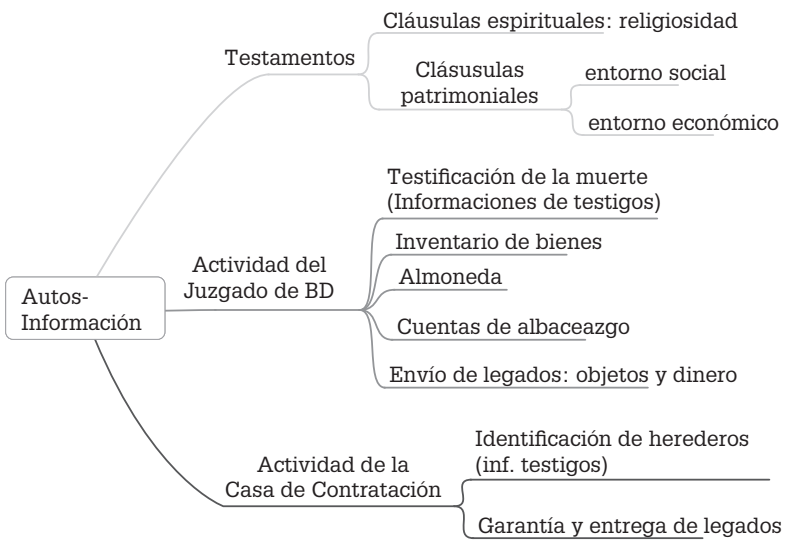
En el descubrimiento, la conquista y la colonización de América, el papel de los particulares tuvo una importancia extraordinaria. La intención de este proyecto es buscar medios para recuperar ese protagonismo que los aspectos más oficiales de la historia tradicional suelen dejan marginado. El problema fundamental que plantea este objetivo es conseguir las fuentes adecuadas que permitan entrar en lo que podríamos considerar el mundo de la gente corriente, es decir, los pobladores españoles que se trasladaron, se asentaron y se mezclaron con otros grupos para construir el mundo colonial identificándose con una América nueva. Esas fuentes tienen que ser documentos privados que se hayan conservado a través de un proceso de oficialización. En este caso son, sobre todo, testamentos y cartas que se encuentran en el Archivo General de Indias de Sevilla, en

[6] García-Abásolo, *La música de las catedrales*.

[7] Celebrado en la Escuela de Estudios Hispano-Americanos (CSIC), Sevilla, febrero de 2010.

un apartado singular de la Sección de Contratación en el que se recogen los autos de bienes de difuntos. Estos documentos pasaron a la administración colonial a causa de la responsabilidad asumida por la Corona de hacer respetar la voluntad de los testadores y los derechos de sus herederos. De tal manera esos documentos que están ligados a la muerte de los pobladores se convierten en la fuente fundamental para poder entrar en su vida privada y devolverles su protagonismo en la construcción humana, social, económica y cultural de la América española. Al cuerpo principal de testamentos conservados en el Archivo General de Indias le hemos añadido otros del Archivo de Protocolos de Córdoba, del Archivo General del Obispado de Córdoba y del de Notarías de México.[8]

1. Campos de información posible de los autos de bienes de difuntos



Campos de información de los testamentos de los pobladores de Indias

Queremos mostrar que, mediante el análisis informático, el historiador puede disponer de un gran acervo de datos en cada una de las partes de que suelen constar los testamentos y, además, de la capacidad de relacionar esos datos entre sí y con otros afines. El resultado final es la estruc-

[8] Catálogo de protocolos del Archivo General de Notarías.

turación del testamento de manera que pueda ser analizado en cuanto a la información que contiene en sí mismo y en su relación con la de los demás que constituyen nuestra base de trabajo.

Aprovechando la regularidad del testamento como documento, podemos intentar un desarrollo sistemático. A través del testamento disponemos de información en los apartados siguientes:

Información cultural y de los sentimientos religiosos

Dentro de este primer campo englobaríamos asuntos relacionados con tradiciones y devociones, pertenencia a cofradías y otras asociaciones piadosas, orientación hacia un campo determinado de espiritualidad (franciscanos, dominicos, agustinos, jesuitas). En este punto, las mandas para celebrar misas por el alma de los testadores y los legados a los conventos, aunque en cada caso estaban en relación con la fe, la generosidad y las posibilidades económicas de los testadores, consideradas en conjunto llegaron en ocasiones a significar cantidades nada despreciables de recursos para las personas y entidades beneficiarias. Un ejemplo tomado de 24 testamentos de vecinos españoles de Manila de finales de siglo XVI y principios del XVII será suficientemente explícito de esto: considerando sólo las mandas de misas cuantificables –que no fueron las únicas–, la cantidad recibida por las iglesias y los conventos de Manila, con preferencia de los de San Agustín, San Francisco, Santo Domingo y de la catedral, fue de más de 26 000 pesos.[9]

Dentro de este mismo campo señalaríamos la valoración de las formas de religiosidad popular y de las tradiciones llevadas por los pobladores españoles a América, y de las americanas que hicieron el camino inverso; valoración de la información relativa al arraigo en América de los pobladores españoles (criollización), y a lo que podíamos llamar americanización de los lugares de origen de esos pobladores por los contactos regulares mantenidos con sus parientes.[10] Estos aspectos tienen un complemento de enorme interés a través de los inventarios de bienes y de las almonedas que se hacían para transformarlos en dinero y cumplir los legados testamentarios. Además de los utensilios domésticos y el ajuar de

[9] García-Abásolo, "Aplicación de técnicas", pp. 311-335.

[10] En este aspecto deben tenerse en cuenta como documentación complementaria los protocolos notariales y las cartas cruzadas entre los pobladores de los reinos de ultramar y sus familias en España. Muchas de estas cartas se conservan en el Archivo General de Indias y una buena parte ha sido publicada en algunos compendios específicos. Lockhart y Otte, *Letters and People*; Otte y Albi, *Cartas privadas de emigrantes*. Una muestra de la utilidad de estas cartas en Solano, "Elites y calidad de vida", pp. 139-162.

ropa, contienen objetos relacionados con el nivel y preferencia culturales (libros) y con el mundo de las devociones (imágenes y objetos religiosos), que indican el trasvase cultural y religioso de España a América y de América a España.

Teniendo en cuenta las dificultades de la navegación oceánica, a veces resulta sorprendente la rapidez con la que se trasladaban los libros a lo largo del Imperio. En el testamento de Pedro de Zúñiga, un comerciante vecino de Manila de fines del siglo XVI, figura un conjunto de libros de resto de las partidas que solía negociar con Albarránez Freyre, su proveedor en el virreinato de Nueva España, avecindado en Puebla de los Ángeles. Entre los libros que tenía Pedro de Zúñiga cuando murió había uno con motetes de Francisco Guerrero, afamado maestro de capilla de la Catedral de Sevilla, que pueden ser copias de los editados en Venecia en 1570 y 1589.[11] El testamento está fechado en Manila, el 10 de diciembre de 1607, y en él figuran algunas partidas de libros entre los que estaban los motetes referidos (cinco libros), seis pasionarios de canto llano, un juego de motetes de madrigal (cuatro libros), siete libros de Semana Santa y siete breviarios. Son interesantes los libros de música y también los demás, en la medida en que reflejan la llegada a Manila de contenidos litúrgicos y rituales desde Nueva España.[12]

Información sobre la actividad profesional

Testamentos, inventarios y almonedas de bienes también permiten valorar la significación de actividades profesionales determinadas en las localidades americanas en las que se asentaron los pobladores que conocemos. Esto tiene particular relieve cuando la actividad profesional produce artesanía selecta, como es el caso por ejemplo de los plateros y orfebres. En la almoneda de los bienes de Diego Cornejo, platero de Salamanca y vecino de Santo Domingo (La Española), efectuada el 16 de septiembre de 1571, compraron buena parte de sus bienes cuatro colegas, que posiblemente constituían el gremio completo de los plateros de Santo Domingo en ese tiempo: Gaspar de los Santos, Baltasar de los Reyes, Hernando Pallarés y Pero Ruiz. Es bastante razonable que, en un mercado escasamente abastecido, el material profesional de una cierta

[11] Albarránez Freyre enviaba cajas de libros y Zúñiga le pagaba con mercancías chinas, japonesas y filipinas para vender en el mercado virreinal. Summers, "Listening for historic Manila", p. 203.

[12] AGI, Contratación 287, N1, R15, *Autos sobre los bienes de Pedro de Zúñiga, natural de Torija, en Guadalajara, y fallecido en Manila en 1608*; García-Abásolo, "The Private Environment", pp. 349-373.

especialización fuera muy bien recibido. Por otra parte, las descripciones de las joyas que se vendieron en esa almoneda, así como los precios que alcanzaron, son datos poco habituales y de gran utilidad para la historia de la orfebrería en Santo Domingo a mediados del siglo XVI.[13]

Esperamos dar continuidad a nuestro proyecto aplicando las mismas técnicas y metodología a los inventarios y almonedas de bienes de los pobladores de las Indias españolas. Los testamentos de los indios están siendo fundamentales para poder estudiar el proceso de transformación de la agricultura prehispánica en la época virreinal. Se han encontrado datos en estos testamentos de indígenas sobre las labores agrícolas y ganaderas que emprendieron, la tecnología y el utillaje que emplearon y el entorno económico y social al que pertenecieron. Son documentos escritos en náhuatl y en castellano en los que hay noticias detalladas de tierras, animales, plantas y aperos que usaron los indígenas en los siglos XVI y XVII, de manera que constituyen un documento básico para el estudio de la agricultura en la época colonial. El campo va más allá de la información sobre las transformaciones agrícolas.[14]

El mundo social y económico de los pobladores

Una puerta de entrada a esta temática podría ser la valoración de la capacidad profesional y del volumen de negocio, que en casos precisos, como los de comerciantes, mineros y hacendados, es un complemento valioso y singular para esos aspectos de la historiografía tradicional. Lo mismo puede decirse de las redes sociales y económicas integradas por familiares, parientes y paisanos que muestran la importancia de las vinculaciones familiares y locales en la colonización de América. En ocasiones, los expedientes administrativos generados con motivo del fallecimiento de un poblador son la única vía posible para conocer aspectos del mundo colonial español que permanecían ocultos. Por ejemplo, a través del testamento del filipino Domingo de Villalobos, comerciante que recorría con sus mulas una amplia extensión de la costa del Pacífico mexicano, se puede ver en funcionamiento una comunidad de casi veinte filipinos, asentados en pueblos de indios y en huertas de cacao, con los que Domin-

[13] AGI, Contratación 209, N. 1, R. 1, *Almoneda de los bienes de Diego Cornejo, Santo Domingo, 16 de septiembre de 1571. Autos sobre los bienes de Diego Cornejo, natural de Salamanca y muerto en Santo Domingo*; García-Abásolo, "El mundo privado", p. 22; García-Abásolo, *La vida y la muerte*.

[14] Rojas, Rea y Medina, *Vidas y bienes olvidados*; Ruz, "De antepasados y herederos", pp. 7-32. Rodríguez, *Testamentos indígenas*.

go de Villalobos trataba habitualmente.[15] El trabajo continuo sobre estos documentos está enriqueciendo nuestro conocimiento del mundo colonial desde nuevas perspectivas que permiten llegar desde lo más material y cotidiano, hasta lo más espiritual e íntimo.[16]

La repercusión de la actividad de los pobladores de Indias en sus lugares de origen.

Es posible, asimismo, valorar el éxito o el fracaso de la experiencia indiana a través de los legados testamentarios que, en forma de efectivo para invertir a favor de familiares y de fundaciones, beneficiaron también a un gran número de paisanos que suscribieron préstamos en partidas normalmente pequeñas y muy distribuidas a un interés moderado. La documentación de estas fundaciones muestra que los beneficiarios más numerosos fueron labradores, pequeños propietarios que pudieron acudir a una especie de microcréditos singulares en una época de escasa liquidez en España. Es posible que éste haya sido el empleo de plata americana que tuvo más alcance social.[17]

Estos legados testamentarios motivados por el buen recuerdo de la tierra de origen consistieron también con frecuencia en objetos, casi siempre de carácter religioso, destinados a las iglesias o ermitas de los lugares de los que los testadores eran oriundos. Cuadros, imágenes, ornamentos (lámparas, coronas y vasos sagrados de oro y plata) pasaron de América y Filipinas a España y en muchos casos todavía se conservan y forman parte de un rico patrimonio que por esta vía se puede conocer en sus pormenores.[18]

15 AGI, Contratación 520, N. 2, R. 14, *Autos y diligencias en razón de la cobranza de los bienes de Domingo de Villalobos, difunto*; García-Abásolo, "Filipinos on the Mexican Pacific".

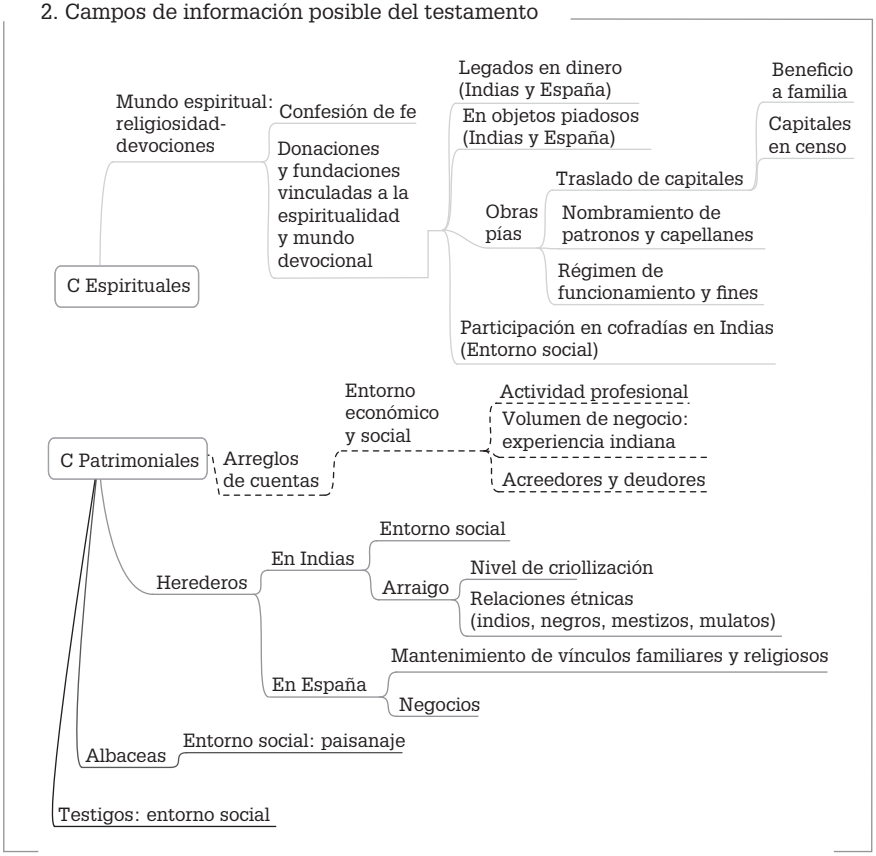
16 Gabbi y Martín de Codoni, *Mendoza en sus testamentos*. Kordic y Goic, *Testamentos coloniales*.

17 Un ejemplo de la importancia y variedad de estas inversiones se puede ver en García-Abásolo, "Los beneficios de tener indianos".

18 El seguimiento de estos aspectos hace aconsejable una selección de pobladores por origen, de manera que se puedan establecer relaciones entre los documentos relativos a los pobladores en América y su parentela en sus localidades de nacimiento. Por eso hemos introducido documentación de los archivos de protocolos y eclesiásticos, de manera que el proyecto aporta los beneficios derivados de la relación complementaria entre los archivos generales, en este caso el General de Indias, con los locales. García-Abásolo, "Inversiones indianas en Córdoba", pp. 427-454; también García-Abásolo, "Los beneficios de tener indianos"; Ortiz de la Tabla, "Emigración a Indias", pp. 441-460.

Los testamentos tienen referencias precisas a muchos pobladores que con frecuencia no se registraron en los libros de pasajeros a Indias de la Casa de Contratación. Por tanto, el aporte de los testamentos es muy importante para el estudio de la emigración real a América. En los testamentos y en las almonedas de venta de los bienes de los testadores figuran muchas personas como parientes, acreedores, deudores, albaceas, testigos, amigos, compradores de los objetos del difunto, que son pobladores localizados en un lugar y en un tiempo determinados. En el apartado que dedicamos a exponer parte de los resultados obtenidos del análisis de la base de datos que hemos utilizado para desarrollar nuestro proyecto ofrecemos algunas cifras que permiten valorar mejor las posibilidades en este ámbito. Los datos sobre pobladores localizados por este medio son una fuente de primer orden para completar los que ya conocemos sobre la emigración a las Indias españolas.

2. Campos de información posible del testamento



Preferencia por los pobladores andaluces

Algunos trabajos sobre la emigración extremeña a América en el siglo XVI, o sobre pobladores cordobeses de los reinos de ultramar entre los siglos XVI y XVIII, muestran la utilidad de acotar en el espacio y en el tiempo los objetivos hasta aquí descritos. Ida Altman se ha enfocado en los emigrantes de Cáceres y de Trujillo que, además, se localizaran preferentemente en el Perú; su trabajo se convierte en modelo útil para establecer las pautas de relación entre una zona española y otra americana en el periodo colonial.[19]

Se han empleado criterios similares para estudiar los andaluces en Indias, en particular los habitantes del antiguo Reino de Córdoba, aunque en periodos más amplios y teniendo en cuenta que la emigración cordobesa fue muy dispersa y se repartió por todo el Imperio, tanto en América como en las Filipinas.[20] Ambos tipos de estudios pueden considerarse una nueva forma de tratar la emigración española a Indias, teniendo en cuenta aspectos sociales, económicos y culturales que pasan inadvertidos cuando las perspectivas se concentran en aspectos cuantitativos y estadísticos.

En nuestro proyecto hemos mantenido el protagonismo de los cordobeses, pero hemos introducido a otros pobladores originarios de distintos lugares de Andalucía y de España e incluso a algunos extranjeros. Al final, la base de datos para analizar está compuesta por 350 testamentos procedentes del Archivo General de Indias, del Archivo Histórico Provincial de Córdoba, del Archivo del Obispado de Córdoba y de algunas notarías de la ciudad de México fechados en los siglos XVI, XVII y XVIII, con preferencia de los siglos XVI y XVII. La mayor parte corresponde al periodo entre 1550 y 1650, una época en la que se pueden valorar mejor los aportes del trasvase humano y cultural y en la que la emigración familiar aparece consolidada, de manera que la actividad de las redes de parentesco y paisanaje en la colonización es más palpable.

19 Aunque Cáceres y Trujillo no representan más que una parte de Extremadura, centrarse en la emigración que salió de esos dos núcleos urbanos aporta las ventajas de facilitar el seguimiento de familias de emigrantes en un tipo de estudio de la emigración con una perspectiva más enfocada en lo social que en lo estadístico. Detrás de estos movimientos aparece el entramado de actividades económicas puesto en marcha por estas redes de familia y paisanaje, y el bagaje cultural de tradiciones y devociones que llevaron consigo. Altman, *Emigrantes y sociedad*.

20 García-Abásolo, *La vida y la muerte en Indias*; García-Abásolo, "Andaluces de Cuba", pp. 55-152. García-Abásolo, Quiles y Fernández, *Aportes humanos, culturales y artísticos*.

Técnicas informáticas historiográficas. Evolución histórica

Los historiadores han venido usando diversas técnicas para el análisis de las fuentes documentales históricas. Con la llegada de la informática muchas de estas técnicas han sido revisadas, incorporando los avances producidos en esta nueva disciplina. En el ámbito de la investigación histórica, las tareas[21] que forman parte de dicho quehacer precisan que las fuentes documentales disponibles sean procesadas en algún formato que sea útil para la recuperación y extracción de la información. Tradicionalmente, hasta la llegada de informática, la información obtenida de las fuentes bibliográficas, documentales y de otro tipo se registraba en fichas de texto expresadas en lenguaje natural. Se les asignaban una serie de índices y se procedía a clasificarlas temáticamente para facilitar su recuperación, o con la intención de elaborar estructuras intermedias de síntesis para la redacción del trabajo final.[22] La informática permite sustituir las fichas en papel por documentos electrónicos organizados de diferentes maneras; para destacar las características de aquellas que se citarán a continuación, imaginemos este fragmento de un documento sobre el que vamos a ver diferentes formas de representación:

> Sepan cuantos esta carta de testamento y última voluntad vieren cómo yo, Eugenio de Chaves Calizares, natural de la villa de Yepes, en el Reino de Toledo, en España, hijo legítimo de Diego de Chaves y de Magdalena de Velasco, su mujer, mis padres, vecinos que fueron de la dicha villa, ya difuntos...

Registro de documentos en archivos de texto

En esta opción la información se registra en documentos escritos en lenguaje natural de forma no estructurada, usando programas convencionales. Algunos investigadores tienden a registrar en un archivo electrónico uno o varios documentos de las fuentes. Los documentos pueden ser registrados en su totalidad o bien un resumen de ellos. Para la obtención de

[21] Hablamos del planteamiento del problema o los problemas históricos a tratar, estudio de los antecedentes del problema, identificación de las fuentes documentales, identificación de las fuentes bibliográficas, elaboración del primer diseño de la estructura del trabajo, recopilación y registro de las fuentes, reelaboración y obtención de la estructura final del trabajo, obtención de estructuras de datos intermedias de síntesis, redacción final del trabajo.

[22] Bernardo y Calvo, *Historia e informática*, pp. 170-185.

estructuras de síntesis o para la elaboración del trabajo el investigador usa técnicas básicas de recuperación de información mediante opciones de búsqueda sencilla que traen incorporados estos sistemas. También pueden usarse sistemas de recuperación de información que incorporen el indexado previo de archivos. La principal ventaja de este sistema es su sencillez, pero su mayor inconveniente es su falta de eficacia.

Otra posibilidad es registrar la información en campos de texto de una tabla sencilla en una base de datos y utilizar los recursos de búsqueda que ofrecen los gestores de bases de datos para realizar la búsqueda y selección de información.

Bases de datos relacionales

Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados de forma estructurada para su posterior uso. En la actualidad, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital (electrónico), que ofrece un amplio rango de soluciones al problema de almacenar datos. Existen programas denominados sistemas gestores de bases de datos que permiten almacenar y posteriormente acceder a los datos de forma rápida y estructurada.[23]

Una base de datos relacional es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permite establecer interconexiones (relaciones) entre los datos (que están guardados en tablas) y a través de dichas conexiones relacionar los datos de ambas tablas, de ahí proviene su nombre de modelo relacional. Para facilitar la creación de una base de datos relacional se propone crear un modelo conceptual basado en la identificación de tipos de entidades y tipos de interrelaciones acompañadas de su representación gráfica.[24] Las entidades llevan atributos para su descripción. Un atributo *id* identifica de manera inequívoca a cada entidad. Este modelo conceptual, fácil de interpretar por una persona no experta en informática puede representarse de manera más próxima al computador a través de un modelo relacional. Tanto las entidades como las relaciones se representan mediante tablas. La fila de una tabla representa un ejemplo del tipo de entidad o del tipo de relación que

[23] Estas representaciones fueron estudiadas en el terreno de la investigación histórica por Thaller, "Methods and Techniques", pp. 147-156. Un estudio comparativo puede encontrarse en Denley, "Models, Source and Users", pp. 33-43.

[24] Es el denominado modelo entidad-relación de Peter Chen. Puede verse una descripción sencilla en http://es.wikipedia.org/wiki/Modelo_entidad-relación.

caracteriza. La sola representación de la información del pequeño extracto del documento transcrito anteriormente requiere de un conjunto de tablas relacionales para representar los hombres, las mujeres, los lugares, las relaciones de parentesco, etcétera.

La gran cantidad de tipos de entidades y tipos de relaciones que aparecen en la enorme variedad de documentos que se utilizan en la investigación histórica hace muy complejo pensar en obtener modelos conceptuales y relacionales capaces de soportar la variada información que se registra. Pensemos en que nada más de fuentes notariales, por ejemplo, documentos de compraventa, imposición y redención de censos, arrendamientos, testamentos, etc., supondrían una gran complejidad.

Trasladar la información desde los textos no estructurados de los documentos a las tablas relacionales de una base de datos supone además un esfuerzo sobreañadido de transcripción, colocando cada uno de los elementos del texto en su celda correspondiente.

Sistemas de información basados en documentos xml

Un intento de superar estos problemas es representar los documentos en formato *xml*. La ilustración 3 muestra el fragmento de texto en este formato *xml*. Serrano Tenllado realizó un trabajo de modelado de los diferentes tipos de documentos utilizando modelos que proporciona *xml-schema*. Obtuvo un alto nivel de reutilización y propuso el registro de la información en un

3. Representación de la información en xml

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<texto>Sepan cuantos esta carta de testamento y última voluntad  vieren como yo,
<hombre id="Eugenio_de_Chaves_Calizares">Eugenio de Chaves Calizares</hombre>,
natural  de la
<lugar id="villa_de_Yepes">villa de Yepes</lugar>, en el <lugar id="Reino_de_Toledo">
Reino_de_Toledo</lugar>, en  <lugar id="Espaya">España</lugar>, <hijo_legitimo_de>hijo
legítimo de<hombre id="Diego_de_Chaves">Diego de Chaves</hombre> y de<mujer
id="Magdalena_de_Velasco">Magdalena de Velasco</mujer></hijo_legitimo_de>, su
mujer, mis padres, vecinos que fueron de la dicha villa, ya difuntos.</texto>
```

sistema semiestructurado desde el cual resultaba bastante eficiente su recuperación.[25] Aunque el sistema es muy eficiente si se lleva a sus últimas consecuencias, la identificación y el marcado de todas las entidades y relaciones de la documentación suponen un esfuerzo de transcripción que no todos los investigadores están dispuestos a realizar. Además quedan hechos que no han podido reflejarse de manera explícita como que *Diego*

[25] Serrano Tenllado, *El poder socioeconómico*. Una descripción básica de xml puede encontrarse en http://es.wikipedia.org/wiki/Extensible_Markup_Language.

de Chaves y *Magdalena de Velasco* fueron un matrimonio legítimo y que en el momento de otorgarse el documento eran difuntos. Tampoco se ha expresado explícitamente mediante elementos marcados que las tres personas mencionadas eran naturales de la villa de Yepes.

Sistemas de información basados en documentos rdf-owl

Una alternativa al paradigma anterior es utilizar modelos de representación basados en la *web semántica*. El objetivo es obtener un modelo del dominio caracterizado por una jerarquía de entidades y relaciones, junto a un conjunto de axiomas que representan una visión de la realidad simplificada, pero al mismo tiempo rica semánticamente. La ventaja de esta representación es que no sólo es comprensible para los agentes humanos sino también para los agentes *software*.

La web semántica es la "Web de los datos". Se basa en la idea de añadir metadatos semánticos y ontológicos a la World Wide Web. Esa información adicional que describe el contenido, el significado y la relación de los datos se debe proporcionar de manera formal, para que así sea posible evaluarla automáticamente por máquinas de procesamiento. El objetivo es mejorar internet ampliando la interoperabilidad de los sistemas informáticos usando "agentes inteligentes". Agentes inteligentes son programas en las computadoras que buscan información sin operadores humanos.

Para ello se han propuesto formalismos de representación (*rdf*, *rdf-schema*, *owl*) de la información y el conocimiento que hacen posible no sólo la recuperación de la información mediante consultas (*sparql*), sino además pueden inferir nueva información o detectar inconsistencias mediante la declaración de reglas de inferencia (*swrl*) y el uso de programas razonadores (*pellet*, *kermit*, etcétera).[26]

Estas representaciones muestran una visión de la realidad a través de una colección de tripletas de la forma (sujeto, predicado, objeto). Una parte de la representación se dedica a establecer una jerarquía de clases o categorías semánticas mediante relaciones de tipo *es--un*. Se definen además propiedades de tipo dato (numéricas, de tipo texto, de tipo fecha, etc.) que relacionan las instancias de las clases con los diferentes tipos de datos y de tipo objeto, que relacionan dos instancias de clases entre sí.

El siguiente código ilustra un fragmento de la declaración de las propiedades y clases de esta pequeña ontología que está sirviendo para ilustrar el trabajo. Las instancias de la ontología (los individuos, las cosas y sus relaciones) quedan representadas por las tripletas.

[26] Una descripción de estas tecnologías pueden verse en http://www.w3.org/standards/semanticweb/.

```
<NamedIndividual rdf:about="&ejemplo;Diego_de_Chaves">
<rdf:type rdf:resource="&ejemplo;Hombre"/>
<ejemplo:nombre>Diego de Chaves</ejemplo:nombre>
<ejemplo:esposoDe rdf:resource="&ejemplo;Magdalena_de_Velasco"/>
<ejemplo:natural_de rdf:resource="&ejemplo;villa_de_Yepes"/>
</NamedIndividual>
```

Este tipo de representación tiene varias ventajas sobre las bases de datos relacionales: la declaración de los modelos de representación puede ubicarse en la red de Internet y reutilizarse, de manera que un usuario podría importar la ontología y centrarse en la creación de instancias o ejemplos de las entidades y relaciones descritas en ella. Es como si para usar una base de datos en *Microsoft-Access* se pudieran importar las tablas, los campos de cada tabla, las restricciones a dichos campos y las relaciones entre dichas tablas. Todo ello está configurado en la ontología donde se declaran las clases, las propiedades y sus restricciones. El usuario, al crear una nueva ontología para la representación de la información, importa la ontología donde está declarado todo este conocimiento y se centra en la creación de las instancias.[27]

Las reglas inferenciales también están creadas en la ontología de base. No es necesario crear múltiples tablas interconectadas que expresen diversas relaciones. Con este modelo, lo que se crea es un gran grafo con múltiples nodos relacionados entre sí. El lenguaje de recuperación de información, *sparql*, es intuitivo y próximo a lenguaje *sql*.[28]

Sistemas basados en procesamiento de lenguaje natural

Independientemente del formalismo que se elija para la representación de la información (documentos de texto no estructurado, bases de datos relacionales, documentos basados en la tecnología *xml*, o documentos basados en ontologías *rdf-owl*) es necesario trasladar la información desde las fuentes documentales originales a estos formalismos de representación.

En la situación actual, si las fuentes documentales que se manejan son fuentes históricas manuscritas, hay que realizar la transcripción hacia estas otras representaciones electrónicas. Lo más sencillo es hacerlo sobre documentos de texto no estructurado y, posteriormente, desde ahí trasladar lo que interese a otro formalismo. Además, en el futuro es previsible que haya gran cantidad de información no estructurada en formato

[27] El término instancia se utiliza aquí en el sentido que se le da en informática como un ejemplo particular de una clase.

[28] Ver http://www.w3.org/TR/rdf-sparql-query/.

electrónico y sería útil contar con técnicas que permitan pasar de forma automática la información de una a otra representación más formal, desde donde puedan obtenerse estructuras de información de síntesis útiles para la elaboración del trabajo final.

Por todo ello, pensamos que es de gran utilidad contar con un sistema de procesamiento automático de lenguaje natural que permita extraer información de modo automático y representarla en algunos de los formalismos previamente mencionados.

A continuación analizaremos qué tareas son necesarias para la construcción de un sistema que haga posible esto. Se estudiará la arquitectura del sistema identificando los diferentes subsistemas, los módulos de que se compone cada uno de estos subsistemas y las relaciones que existen entre ellos. Se abordará su eficacia y fiabilidad, y, finalmente, se expondrán las conclusiones y los futuros retos a los que nos enfrentaremos.

Identificación del conocimiento

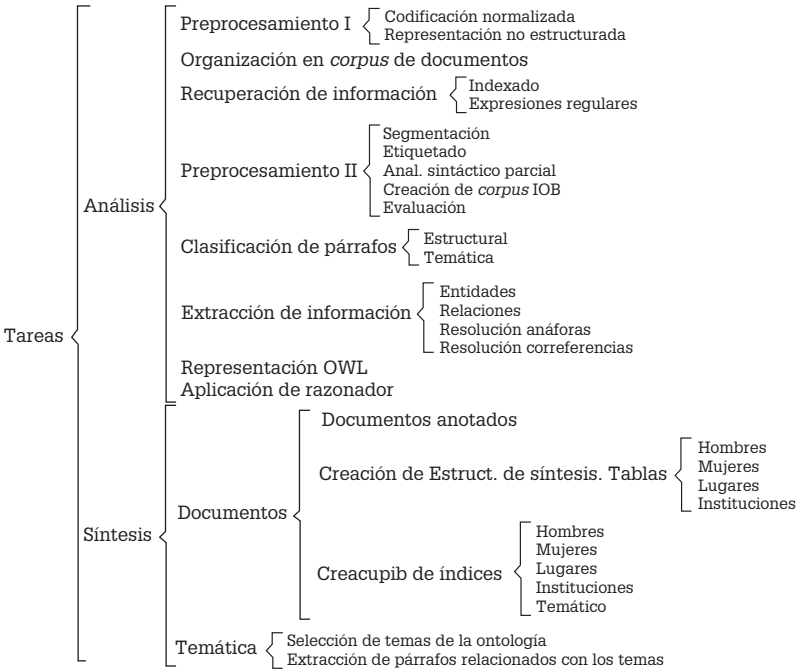
En este trabajo partimos de la premisa de que toda la información está registrada en documentos electrónicos no estructurados. El objetivo es obtener un conjunto de estructuras y documentos de síntesis que faciliten la labor del investigador y mejoren su rendimiento y la calidad de los trabajos que realice. Para nuestro experimento partimos de un conjunto de 330 testamentos de pobladores de América y de Filipinas. La identificación del conocimiento la haremos siguiendo las indicaciones de la metodología *CommonKads*.[29]

Conocimiento de las tareas

La ilustración 4 muestra las tareas que se han identificado para el sistema que se propone construir. Las hemos dividido en tareas de análisis y tareas de síntesis. Las primeras comprenden las actividades necesarias para el acopio, la organización, la representación, la clasificación y la extracción automática de la información. Las segundas comprenden las actividades necesarias para la obtención de información de síntesis del sistema útil para la elaboración de los trabajos finales de investigación. En la siguiente sección describiremos las tareas identificadas más importantes.

[29] Un estudio amplio de esta metodología puede verse en Schreiber, *Knowledge Engineering and Management*.

4. Identificación de tareas



Conocimiento del dominio. Ontologías del dominio

Entendemos por dominio un área de trabajo especializada en la cual van a realizarse un conjunto de tareas. Así, podemos hablar del dominio de la oftalmología dentro del área de la medicina, o siendo más específicos, del dominio del glaucoma como un subdominio de la oftalmología. En el campo del procesamiento del lenguaje natural la tarea de extracción de información podrá aplicarse a diferentes dominios, por ejemplo abordar el análisis de historias clínicas en el dominio del glaucoma. También podría aplicarse esta tarea en el campo del derecho en el dominio del fraude fiscal.[30]

[30] Recogemos aquí la propuesta que subyace en la metodología CommonKads, la reutilización. Se propone que las tareas y sus diferentes métodos de resolución son genéricos, susceptibles de aplicarse en diferentes dominios. De igual forma la caracterización de un dominio representada formalmente a través de su ontología del dominio podrá ser utilizado para resolver diferentes tareas.

Todo dominio se caracteriza por un vocabulario especializado que describe tipos de entidades y tipos de relaciones entre estos tipos de entidades. Cada tipo de entidad o relación puede tener propiedades que satisfagan determinadas restricciones. Incluso pueden identificarse reglas que permitan realizar inferencias sobre determinadas entidades u obtener nuevas relaciones.[31]

Si observamos el texto que nos viene sirviendo de ilustración, el dominio puede situarse en el campo de la historia social. Pueden identificarse claramente tipos de entidades (clases) como personas, hombres, mujeres, lugares, y entre ellas establecer relaciones familiares, toponímicas, etc. Del texto pueden inferirse nuevas relaciones familiares que no están explícitamente descritas en él. Gracias al conocimiento acumulado por el ser humano, al leer el texto en lenguaje natural resulta muy simple deducir que *Diego de Chaves y Magdalena de Velasco* estaban casados legítimamente. Desgraciadamente, las máquinas no poseen ese conocimiento y hemos de declararlo formal y explícitamente.

En el dominio de la historia, el investigador, ante un problema, conoce qué fuentes debe de consultar, qué entidades y relaciones tienen interés para su estudio o cómo va a estructurar su trabajo temáticamente. Para cada uno de los epígrafes del trabajo sabe qué palabras clave debe buscar en las fuentes de manera que pueda seleccionar los párrafos que sean más útiles para su redacción. Sabrá además qué estructuras de síntesis le serán necesarias, entre otras cosas.

Todo este conocimiento del investigador habrá que identificarlo y representarlo formalmente en una ontología para que agentes *software* puedan realizar por él algunas tareas, de forma que se incremente la eficacia de su trabajo. No debemos caer en la ingenuidad de pensar que estos agentes serán capaces de escribir el trabajo por el investigador, pues es evidente que no poseen el extensísimo conocimiento especializado que tiene el ser humano; no tienen sentido común ni saben cómo tratar el conocimiento tácito y, por supuesto, no disponen de sus destrezas.

Sin embargo, cuando se delimitan las tareas a realizar y el conocimiento del dominio que se va a manejar, es posible que los agentes *software* puedan realizar actividades de manera mucho más eficiente que el ser humano, aunque posiblemente de forma menos precisa. Supongamos la necesidad de analizar miles de documentos de una determinada naturaleza, por ejemplo testamentos expresados en lenguaje natural. Lo que a un investigador le podría llevar meses un agente *software* lo podría hacer

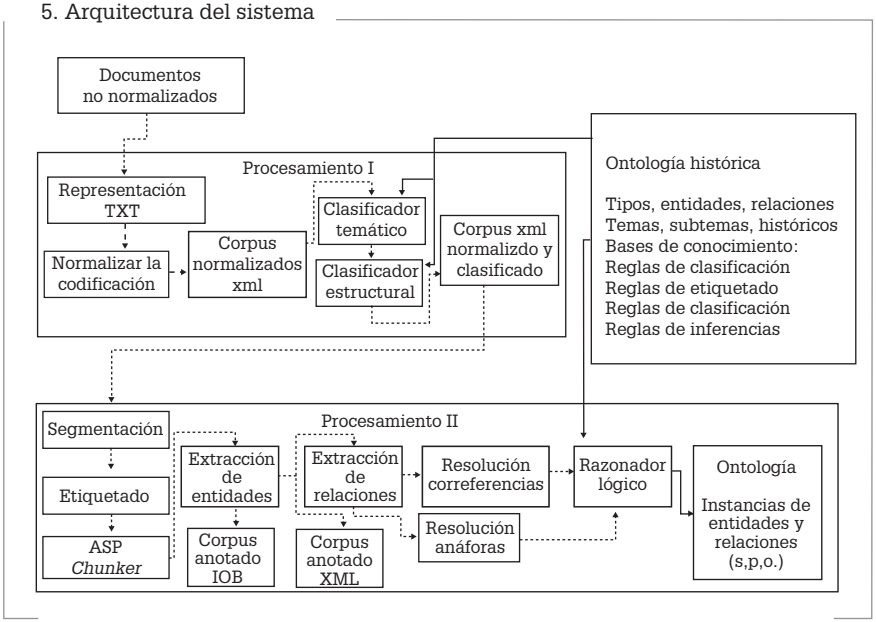
[31] Una revisión sobre el término ontología en el dominio de la investigación histórica puede verse en Serrano Tenllado, *El poder socioeconómico*, pp. 39-41.

en horas. Todas estas consideraciones son las que se recogen en una ontología del dominio. Afortunadamente hoy se dispone de formalismos y herramientas para construir estas ontologías.[32]

Arquitectura del sistema propuesto

Para el desarrollo de la aplicación se ha utilizado como arquitectura de referencia la arquitectura Modelo de la aplicación, Vista, Controlador (MVC).[33] En este trabajo sólo abordaremos el componente del modelo de la aplicación, describiendo las principales tareas que se llevan a cabo, así como el conocimiento del dominio. Las tareas corresponderán a procesos que manipulan elementos del conocimiento del dominio como las fuentes documentales, hacen uso de bases de conocimiento del dominio y dan como resultado otros elementos del dominio como tablas, documentos, etc., que representan *estructuras de síntesis* o *documentos de síntesis*.

5. Arquitectura del sistema



[32] Una de las herramientas más extendidas es PROTÉGÉ cuya descripción puede verse en http://protege.stanford.edu/

[33] http://es.wikipedia.org/wiki/Modelo_Vista_Controlador.

Preprocesamiento I: normalización y clasificación de párrafos

La ilustración 5 muestra parte de la arquitectura del modelo de la aplicación dentro de la arquitectura global del sistema. Un conjunto de documentos no normalizados es sometido a un primer preprocesamiento, normalizando la codificación de los caracteres y buscando una representación en forma de texto uniforme. La salida de este preprocesamiento es un *corpus* de documentos *xml* donde se ha definido un elemento *texto* y dentro de él se ha definido un conjunto de elementos *xml* denominados *p*, que representan párrafos del texto.

Estos párrafos del texto se clasifican temática y estructuralmente a partir de un conocimiento que ha sido declarado en la ontología general a la que hemos llamado *ontología histórica*. Esta clasificación temática queda reflejada en los atributos del elemento *p*. Se han utilizado varios métodos para llevar a cabo esta clasificación (métodos intensivos en conocimiento –método de la poda– y métodos estadísticos –bayesianos, basados en árboles de decisión). Disponer de un *corpus* de documentos con los párrafos del texto clasificados facilita la labor de recuperación de información.

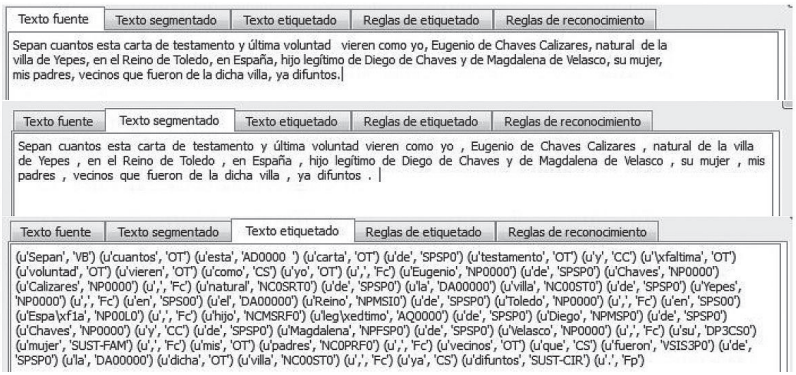
Preprocesamiento II: extracción automática de nombres de entidades y de relaciones

Un segundo módulo de preprocesamiento (ver ilustración 5) está dedicado a la extracción de nombres de entidades. El objetivo es que a partir del conjunto de documentos normalizados y clasificados se pueda identificar y extraer de forma automática un conjunto de entidades y relaciones de interés para el investigador, de forma que permita crear en un nuevo *documento de síntesis* una representación de las fuentes documentales donde, además de incluir los propios documentos, podamos también incluir *estructuras de síntesis* (tablas de hombres, mujeres, instituciones y lugares), índices onomástico, toponímico y temático. El salto cualitativo es importante pues contar con este documento de síntesis facilita la localización de todas estas entidades. Lograr este objetivo no es fácil, pues esta tarea incluye la resolución de problemas de correferencia y anáforas en el texto mediante técnicas de procesamiento de lenguaje natural. Este preprocesamiento contiene las subtareas de segmentación, etiquetado, análisis sintáctico parcial (ASP), extracción de entidades, extracción de relaciones, aplicación de razonadores, anotación de entidades.[34]

[34] Una revisión sobre este problema pueden encontrarse en Padró, "Tendencias en el reconocimiento de entidades", pp. 37-57; Jurafsky, *Speech and Language Processing,* pp. 759-798.

La tarea de *segmentación* consiste en identificar y aislar todos los elementos léxicos del texto. La ilustración 6 muestra el texto original y debajo el texto segmentado. Puede observarse cómo se han aislado los signos de puntuación de las palabras del texto. A cada uno de estos componentes léxicos los denominaremos *token*.

6. Segmentación y etiquetado



El *etiquetado* del texto consiste en asignar a cada componente léxico obtenido de la tarea anterior una etiqueta que representa la categoría léxica del término en la oración. Las categorías léxicas se corresponden a nombres comunes, nombres propios, adjetivos, artículos determinantes, etc. El analizador morfológico desarrollado para el castellano utiliza un conjunto de etiquetas para representar el aspecto morfológico de las palabras. Este conjunto de etiquetas se basa en las etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y *corpus* para todas las lenguas europeas.[35] Para el etiquetado de los componentes léxicos (*tokens*) hacemos uso de una técnica basada en expresiones regulares.[36] Una expresión regular es un patrón que representa a un conjunto de cadenas de caracteres. Cuando una cadena de caracteres cumple el patrón representado en la expresión regular decimos que dicha cadena casa o encaja con el patrón. Haciendo uso de estos patrones podemos construir un conjunto de reglas de etiquetado para que un proceso algorítmico vaya asignando a cada *token* una etiqueta. La regla de etiquetado se compone de dos partes. En la primera

35 http://www.ilc.cnr.it/EAGLES96/intro.html

36 Bird, Klein y Loper, *Natural Language* , cap. 5.

se especifica el patrón y en la segunda la etiqueta que asignará a los *tokens* que encajen con ese patrón. El proceso algorítmico de etiquetado irá aplicando secuencialmente las reglas hasta que encuentre una que encaje con el *token* que se está procesando. Una vez que el *token* ha sido etiquetado continúa con el siguiente *token*. La última regla la haremos de forma que nos aseguraremos que el *token* queda etiquetado con una etiqueta por defecto que señale que dicho *token* no ha sido identificado. Para facilitar la elaboración de ese conjunto de reglas de etiquetado hemos creado un fichero de texto que contiene las reglas de etiquetado que son fácilmente modificables desde cualquier editor de texto.[37] El programa de etiquetado leerá e interpretará este fichero para crear y aplicar el proceso algorítmico de etiquetado. La tercera parte de la ilustración 6 muestra el resultado de aplicar la tarea de etiquetado al resultado previo de segmentación.

El siguiente paso en la arquitectura propuesta en la ilustración 5 es el *Análisis Sintáctico Parcial* (ASP). En este trabajo hemos utilizado una técnica lingüística basada en gramáticas. Se ha creado un fichero de texto que contiene la gramática. El procesador lee este fichero y construye con él un reconocedor de nombres de entidades que aplica a la lista de tuplas del texto etiquetado, obteniendo un árbol de estructuras del cual se extraen los nombres de las entidades.

La técnica básica que hemos usado para la detección de los nombres de las entidades es la estructuración parcial de las frases gramaticales (*chunking*), que identifica y etiqueta secuencias *multi-token* como se ve en la ilustración 7. Aunque no se realiza un análisis gramatical completo, sí es posible aislar y etiquetar estructuras de texto complejas que contienen los nombres de las entidades que se buscan.[38]

El analizador devuelve un árbol de estructuras. El siguiente paso consistirá en recorrer el árbol e identificar nombres para la *extracción de entidades reconocidas*. En nuestro estudio nos hemos centrado en nombres de instituciones, de hombres, de mujeres y de lugares. La ilustración 7 muestra el árbol que se obtiene y, finalmente, la lista de 2-tuplas indicando el nombre y el tipo de la entidad.

[37] El problema de la desambiguación morfológica es la elección del análisis morfológico correcto para una palabra dentro del contexto de una frase entre todos los análisis morfológicos válidos para esta palabra. Existen varios métodos estadísticos para abordar el problema que conlleva la utilización de corpus de texto etiquetado. En este trabajo no abordaremos esta cuestión.

[38] Una descripción de esta técnica puede verse en Bird, Klein y Loper, *Natural Language*, cap. 7.

7. Árbol de reconocimiento

```
(u'como', 'CS')
(u'yo', 'OT')
(u',', 'Fc')
   HOMBRE
   (u'Eugenio', 'NPMSP0')
   (u'de', 'SPSP0')
   (u'Chaves', 'NP0000')
   (u'Calizares', 'NP0000')
(u',', 'Fc')
   RT
   (u'natural', 'NC0SRT0')
   (u'de', 'SPSP0')
(u'la', 'DA00000')
   LUGAR
   (u'villa', 'NC00ST0')
   (u'de', 'SPSP0')
   (u'Yepes', 'NP0000')
(u',', 'Fc')
(u'en', 'SPS00')
(u'el', 'DA00000')
   INSTITUCION
   (u'Reino', 'NPMSI0')
   (u'de', 'SPSP0')
   (u'Toledo', 'NP0000')
(u',', 'Fc')
(u'en', 'SPS00')                    Eugenio de Chaves Calizares m
   LUGAR                            villa de Yepes l
   (u'Espa\xf1a', 'NP00L0')         Reino de Toledo i
(u',', 'Fc')                        España l
(u'hijo', 'NCMSRF0')                Diego de Chaves m
(u'leg\xedtimo', 'AQ0000')          Magdalena de Velasco w
(u'de', 'SPSP0')
```

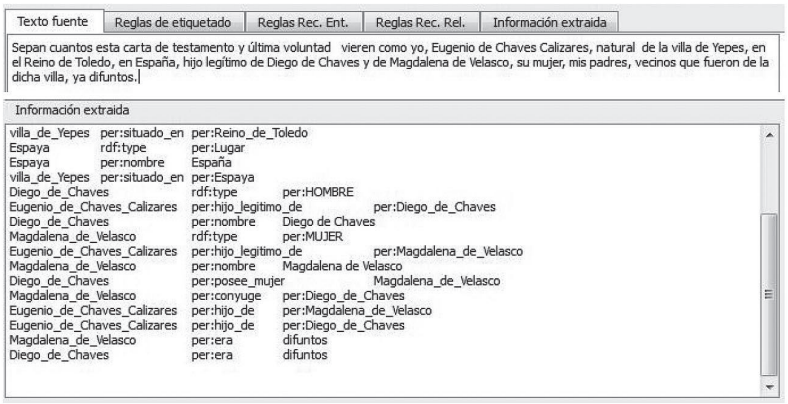
Identificadas las entidades de interés, nuestro objetivo es determinar y *extraer las relaciones* que aparecen entre ellas. La ilustración 8 muestra el texto original y las relaciones que han podido obtenerse de ese texto. Este problema no es trivial, ya que, además de identificar las entidades y sus relaciones entre sí, es necesario resolver problemas de correferencia y de anáforas. Después de la segmentación y etiquetado del texto habría que aplicar dos conjuntos de reglas de reconocimiento: un conjunto para determinar las entidades y otro para determinar las relaciones entre ellas.

Observemos cómo habría que resolver las anáforas "su mujer" y "mis padres" en la frase "*...en España, hijo legítimo de Diego de Chaves y de Magdalena de Velasco, su mujer, mis padres,....*" Las palabras "*su mujer*" hacen referencia a que Diego de Chaves es el marido de Magdalena de Velasco y que Magdalena de Velasco era la legítima mujer de Diego de Chaves. De igual forma las palabras "*mis padres*" hacen referencia a que Diego de Chaves y Magdalena de Velasco eran los padres del sujeto

principal de la oración, Eugenio de Chaves Calizares. Aunque no hemos resuelto en su totalidad este problema, hemos desarrollado varios algoritmos válidos en casos de algunos tipos de anáforas.

La información extraída se muestra en tripletas de la forma (sujeto, predicado, objeto) y de esta forma puede representarse en una ontología mediante el lenguaje *owl*.

8. Extracción automática de entidades y relaciones y aplicación posterior de un razonador lógico



Razonadores. La ilustración 8, además de presentar la información extraída que explícitamente se dice en el texto, muestra información adicional que no se declara en el texto. Por ejemplo, aparece que "*Magdalena_de_Velasco per:conyuge per:Diego_de_Chaves*". Este tipo de afirmaciones pueden obtenerse tras aplicar a la primera información extraída del reconocimiento un conjunto de reglas que pueden estar explícitamente declaradas en el modelo conceptual de la ontología. Existen lenguajes como *swrl* o *CLIPS* que lo hacen posible.

La siguiente etapa en la arquitectura propuesta en la ilustración 5 es la *anotación de entidades* en el documento. El objetivo de esta tarea es modificar el texto de éste, dejando constancia explícita de la parte del texto en la que se ha identificado una entidad. Puesto que la representación del texto de los documentos se ha realizado en *xml*, se han añadido nuevos elementos al modelo del documento de forma que admita la inclusión de estas nuevas estructuras. La ilustración 9 presenta el texto de nuestro ejemplo donde se han reconocido y anotado nombres de entidades.

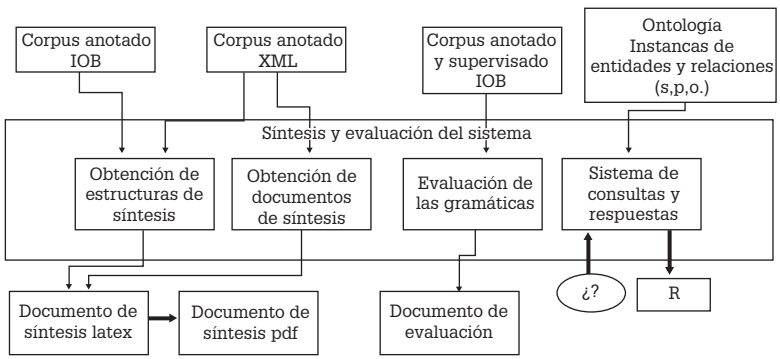
9. Texto anotado

```
<p realThemaCat="" strutctCat="" realStrutctCat="" id="0" themaCat="">En el nombre de Dios
amén. Sepan cuantos esta carta de testamento y última voluntad y pstrimera voluntad vieren como
yo, Eugenio de Chaves Calizares, <rt>natural de</rt> la <l>villa de Yepes</l>, en el Reino de
Toledo, en <l>España</l>, hijo legítmo de <m>Diego de Chaves</m> y de <w>Magdalena de
Velasco</w>, su mujer, mis padres, vecinos que fueron de la dicha villa, ya difuntos.</p>
```

Síntesis y evaluación del sistema

Denominamos *estructuras de síntesis* a aquellas estructuras de datos organizadas que son obtenidas de las fuentes documentales de forma automática y que facilitan al investigador la elaboración del trabajo final. Ejemplos de esas estructuras podrían ser los nombres de las entidades y de las relaciones reconocidas. Denominamos *documentos de síntesis* a aquellos documentos que podemos obtener de forma automática y que incorporan estructuras de síntesis cuya misión es facilitar al investigador la elaboración de un trabajo final. La ilustración 10 muestra las tareas encargadas de realizar la síntesis de estos elementos. A partir de los *corpus* anotados IOB pueden extraerse las entidades reconocidas en las tareas anteriores. Por otra parte, los documentos del *corpus* anotado *xml* han sido sometidos a un conjunto de transformaciones *xslt*, con lo que se ha obtenido de ellos un documento *latex* y posteriormente un documento *pdf* que integra los documentos originales, su localización, las estructuras de síntesis que de ellos han sido obtenidas y un conjunto de índices (onomástico, toponímico, de instituciones, etc.).[39] En este punto, el sistema ha sido sometido a un conjunto de pruebas para validar su comportamiento.

10. Tareas de síntesis y evaluación



[39] Un ejemplo de este documento para un conjunto reducido de documentos puede encontrarse en http://www.uco.es/grupos/aaf/projects/pln/documentos/tmp/-principal.pdf.

Organización de la información

La organización de la información ha mostrado ser eficiente. Desde el punto de vista del usuario no ofrece dificultades, ya que representa la información en archivos independientes, uno por cada documento, y en él se transcribe la información en lenguaje natural sin requerir de ningún conocimiento informático sobre análisis y diseño de bases de datos, *xml*, etc. El proceso de normalización también ha sido correcto, todos los errores han sido subsanados sin un alto costo. La información transcrita por el usuario ha sido trasladada a una representación en *xml*, separando los datos de localización del documento de su contenido. El texto fue separado en párrafos de forma correcta.

Segmentación, etiquetado y reconocimiento

La identificación de componentes léxicos se ha basado en expresiones regulares y el sistema se ha comportado de manera eficiente y ha logrado 100% de precisión.

Para el etiquetado de los componentes léxicos hemos utilizado el conjunto de etiquetas propuestas por el grupo EAGLES para la anotación morfosintáctica de lexicones y *corpus* para todas las lenguas europeas. El método empleado consiste en la aplicación de un conjunto de reglas de etiquetado basado en expresiones regulares. Aunque han surgido algunos problemas de desambiguación, el sistema se ha comportado con una precisión de 87%. Una mayor precisión podría obtenerse mediante el empleo de un *corpus* de documentos etiquetados manualmente y utilizando etiquetadores estadísticos. No obstante, consideramos satisfactorios los resultados obtenidos. Una de las ventajas del método propuesto es independizar el código de la aplicación del fichero que contiene las reglas de etiquetado; esto permite introducir mejoras en el comportamiento del sistema modificando sólo este fichero, sin alterar el código de la aplicación.

Reconocimiento de nombre de entidades

Para evaluar el reconocimiento de entidades se ha creado un *corpus* IOB, donde de forma manual se han identificado todas entidades del *corpus* y se han estudiado diversos parámetros de evaluación.

Sobre una base de datos de 330 testamentos de pobladores de América y de Filipinas hemos obtenido los siguientes resultados de identificación: 6 609 hombres, 1 850 mujeres, 850 instituciones, 2 085 lugares.

Se ha utilizado una métrica que incluye varias medidas de evaluación. Para ello se han definido cuatro parámetros:

TP (Cierto positivo): indica que el término ha sido reconocido como un tipo de entidad y es correcto.
FP (Falso positivo): indica que el término ha sido reconocido como un tipo de entidad, pero es falso.
FN (Falso negativo): indica que el término no ha sido reconocido, pero en realidad sí debería haber sido reconocido.
TN (Cierto negativo): el término no ha sido reconocido, lo que es cierto, ya que no corresponde a una entidad.

Asimismo, se expresa la *exactitud* (*accuracy*): mide el porcentaje de entradas en el conjunto de prueba que son correctamente reconocidas; *precisión:* indica cuántos de los elementos que hemos identificado como entidades realmente lo son: *Precisión=TP / (TP + FP)*, es decir el porcentaje de entidades correctamente reconocidas frente al total de las reconocidas. Un sistema será muy preciso si da muy pocas recuperaciones falsas.

Recubrimiento (*recall*): indica cuántas de las entidades que hemos identificado realmente lo son: *Recall=TP / (TP + FN)*. Es el porcentaje de entidades correctamente reconocidas frente al total de entidades que deberían haberse reconocido de ser un sistema perfecto. Un sistema con un *Recall=1* sería aquel que no dejara ninguna entidad sin reconocer. *F-Measure:* combina la precisión y el recubrimiento para dar una única puntuación; se define como la media armónica de la precisión y el recubrimiento FMeasure=(*2 × Precision × Recall*) / (*Precision + Recall*).

Los resultados obtenidos con la gramática utilizada para el reconocimiento de entidades en un subconjunto de documentos ha sido el siguiente:

<ChunkScoring of 84 chunks>
ChunkParse score:
IOB Accuracy: 94.1%
Precision: 85.1%
Recall: 67.9%
F-Measure: 75.5%
<ChunkScoring of 84 chunks>

Conclusiones y futuros trabajos

Hemos desarrollado un sistema que consideramos eficiente para el tratamiento de fuentes documentales con las siguientes características:

a. La entrada al sistema es un conjunto de documentos escritos en lenguaje natural no estructurado y en formato texto ASCII.

b. La salida al sistema es un documento *pdf* en el que se recogen todos los documentos, las tablas de identificación de nombres de entidades y un conjunto de índices relacionados con dichas entidades. Esto hace que el investigador pueda localizar de forma eficiente las partes del texto donde se ubican dichas entidades.
c. La eficiencia del sistema es alta. Mejoras en los ficheros que contienen las reglas de etiquetado y las reglas de reconocimiento pueden incrementar esta eficiencia.
d. El usuario no tiene que conocer tecnologías informáticas para poder hacer uso del sistema.

El sistema admite varias mejoras; entre otras destacamos:

a. Ampliar y perfeccionar el fichero de reglas de etiquetado.
b. Ampliar y mejorar el fichero de reglas de reconocimiento.
c. Creación de un *corpus* documental de documentos etiquetados correctamente de forma que se mejore la eficiencia del etiquetado y se disminuya el problema de desambiguación.
d. Incorporar un sistema de preguntas y respuestas basado en lenguaje natural que facilite la consulta de la información en el sistema.
e. Creación de un *corpus* de documentos con un análisis sintáctico parcial correcto de forma que se mejore la eficiencia del reconocimiento.

Siglas y referencias

AGI Archivo General de Indias, Sevilla, España.

Bibliografía

Altman, Ida

Emigrantes y sociedad. Extremadura y América en el siglo XVI, Madrid, Alianza Editorial, 1992.

Bernardo Ares, José Manuel de y Antonio Calvo Cuenca

Historia e informática. Metodología interdisciplinar de la investigación histórica, Córdoba, Universidad de Córdoba–Servicio de Publicaciones de la Universidad–Caja Sur, 2005.

Bird, Steven, Ewan Klein, y Edward Loper

Natural Language Processing with Python, Cambridge, O'Reilly Media, 2009.

Catálogo de protocolos del Archivo General de Notarías de la Ciudad de México, México, Universidad Nacional Autónoma de México, 2005.

Denley, Peter
"Models, Source and Users: Historical Database Design in the 1990s", *History and Computing*, Manchester, Manchester University Press, 6/1 (1994), pp. 33-43.
Gabbi, Alicia Virginia y Elvira Martín de Codoni
Mendoza en sus testamentos, siglos XVI, XVII y XVIII, 2 vol., Mendoza, Facultad de Filosofía y Letras de la Universidad de Cuyo, 1996.
García-Abásolo, Antonio
"Inversiones indianas en Córdoba. Capellanías y patronatos como entidades financieras", *Actas de las Segundas Jornadas de Andalucía y América. Andalucía y América en el siglo XVI*, t. I, Sevilla, Escuela de Estudios Hispano-Americanos–Universidad Hispanoamericana Santa María de la Rábida–Excma. Diputación de Huelva–Instituto de Estudios Onubenses, 1983, pp. 427-454.
– *La vida y la muerte en Indias: cordobeses en América. (Siglos XVI-XVIII)*, Córdoba, Monte de Piedad–Caja de Ahorros de Córdoba, 1992.
– "The Private Environment of the Spaniards in the Philippines", *Philippine Studies*, vol. 44 (1996), pp. 349-373.
– "Andaluces de Cuba. Siglos XVI a XVIII", en Jesús Raúl Navarro García (coord.), *Cuba y Andalucía entre las dos orillas*, Sevilla, Consejería de Cultura–Consejo Superior de Investigaciones Científicas–Escuela de Estudios Hispano-Americanos–Asociación Cultural La Otra Andalucía, 2002, pp. 55-152.
– "El mundo privado de los pobladores de la América Española", *Ámbitos*, núm. 16 (2006).
– "Los beneficios de tener indianos. Inversiones de plata americana en la campiña de Córdoba", en Francisco Miguel Espino Jiménez (coord.), *Actas de las VII Jornadas sobre Historia de Montilla*, Montilla, Ayuntamiento de Montilla, 2007.
– "Aplicación de técnicas de inteligencia artificial al estudio de los pobladores de Filipinas", *Archivo Agustiniano*, vol. XCV, núm. 213, (enero-diciembre 2011), pp. 311-335, en http://www.uco.es/aaf/garcia-abasolo/.
– "Filipinos on the Mexican Pacific Coast during the Colonial Period (1570-1630)", en *Into the Frontier; Studies in Spanish Colonial Philippines*, en prensa, Manila, University of Asia and the Pacific.
García-Abásolo, Antonio (coord.)
La música de las catedrales andaluzas y su proyección en América, Córdoba, Servicio de Publicaciones de la Universidad y Caja Sur, 2010.

García-Abásolo, Antonio, Fernando Quiles y María Ángeles Fernández
Aportes humanos, culturales y artísticos de Andalucía en México, siglos XVI-XVIII, Sevilla, Junta de Andalucía–Consejería de Cultura–Consejo Superior de Investigaciones Científicas–Escuela de Estudios Hispano-Americanos–Asociación Cultural La Otra Andalucía, 2006.
Jurafsky, Dan y James H. Martin
Speech and Language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2ª ed., Nueva Jersey, Pearson International, 2009.
Kordic, Raïssa y Cedomil Goic
Testamentos coloniales chilenos, Madrid, Iberoamericana–Vervuert, 2005.
Lockhart, James y Enrique Otte (eds.)
Letters and People of the Spanish Indies. The Sixteenth Century, Cambridge, Cambridge University Press, 1976.
Ortiz de la Tabla Ducasse, J.
"Emigración a Indias y fundación de capellanías en Guadalcanal. Siglos XVI y XVII", en *Andalucía y América en el siglo XVI. Actas de las II Jornadas de Andalucía y América,* t. I, Sevilla, Escuela de Estudios HIspano-Americanos, 1983, pp. 441-460.
Otte, E. y Albi G.
Cartas privadas de emigrantes a Indias, 1540-1616, Sevilla, Junta de Andalucía, 1989.
Padró, Lluis
"Tendencias en el reconocimiento de entidades con nombre propio", en María Antonia Martí y Joaquim Llisterri, *Tecnologías del texto y del habla*, Barcelona, Universidad de Barcelona, 2004, pp. 37-57.
Rojas Rabiela, Teresa, Elsa Leticia Rea López y Constantino Medina Lima
Vidas y bienes olvidados. Testamentos indígenas novohispanos, 3 vol., México, SEP-CONACYT, 1998.
Ruz, Mario Humberto
"De antepasados y herederos: testamentos mayas coloniales", *Alteridades*, 2002, 12 (24), pp. 7-32.
Rodríguez, Pablo (editor)
Testamentos indígenas de Santafé de Bogotá, siglos XVI-XVII, Bogotá, IDCT, 2002.
Serrano Tenllado, María Araceli
El poder socioeconómico y político de una elite local. Los regidores de Lucena en la segunda mitad del siglo XVII, Córdoba, Universidad de Córdoba–Caja Sur, 2004.

Schreiber, G. *et al*.
Knowledge Engineering and Management. The Common KADS Methodology, Cambridge, MIT Press, 1999.
Solano, F.
"Elites y calidad de vida en el Alto Perú a mediados del siglo XVII, según la correspondencia de un noble gaditano", en *Andalucía y América en el siglo XVII. Actas de las III Jornadas de Andalucía y América*, (I), Sevilla, Escuela de Estudios Hispano-Americanos, 1985, pp. 139-162.
Summers, W.
"Listening for historic Manila: music and rejoicing in an international city", *Buhdi: A Journal of Ideas and Culture*, Ateneo de Manila University Press, vol. II, núm. 1, (1998).
Thaller, Manfred
"Methods and Techniques of historical computations", en Peter Denley y Deian Hopkin (ed.), *History and Computing*, Manchester, Manchester University Press, 1987, pp. 147-156.

Páginas electrónicas

http://www.w3.org/standards/semanticweb/.
http://www.w3.org/TR/rdf-sparql-query/.
http://www.ilc.cnr.it/EAGLES96/intro.html
http://www.uco.es/grupos/aaf/projects/pln/documentos/tmp/-principal.pdf.

Recibido el 13 de diciembre de 2010 / Aceptado el 08 de junio de 2011