Revista de Humanidades Digitales

La inteligencia artificial en las Humanidades Digitales: dos experiencias con corpus digitales

Artificial Intelligence in the Digital Humanities: Two Experiences with Digital Corpora

Ana García-Serrano
agarcia@isi.uned.es
ETSI Informática-UNED
https://orcid.org/0000-0003-0975-7205

Antonio Menta Garuz
amenta@invi.uned.es
ETSI Informática-UNED
https://orcid.org/0000-0002-3172-2829

RESUMEN

Este artículo se centra en el análisis de dos investigaciones de diverso signo guiadas por la inteligencia artificial dentro del campo de las HD. El primero es una investigación muy conocida y exitosa de dos lingüistas que resuelven un caso de atribución de autoría a través de la construcción de un corpus digital de 150 obras de 40 novelistas italianos. El segundo es la investigación llevada a cabo en el corpus digital DIMH (*El Dibujante Ingeniero al servicio de la Monarquía Hispánica. Siglos XVI-XVIII*), una evolución de la Colección de mapas, planos y dibujos del Archivo General de Simancas (siglos XVI-XVIII), cuyo objetivo fue desarrollar herramientas de soporte a tareas de anotación semántica, búsqueda de información, extracción de relaciones ocultas en los textos y visualización de los resultados para facilitar la investigación de los historiadores. A través de estos dos ejemplos, este artículo busca mostrar los métodos, procesos y posibilidades de éxito en problemas complejos de investigación en Humanidades resueltos con técnicas de procesamiento del lenguaje natural (PLN) y análisis de datos.

PALABRAS CLAVE: inteligencia artificial, Humanidades Digitales, procesamiento del lenguaje natural, aplicaciones web, análisis de datos.

ABSTRACT

This paper focuses on the analysis of two investigations of different sign guided by artificial intelligence within the field of HD. The first is a well-known and successful research carried out by two linguists solving a case of authorship attribution through the construction of a digital corpus of 150 works written by 40 Italian novelists. The second one is the research carried out on the digital corpus DIMH (*El Dibujante Ingeniero al servicio de la Monarquía Hispánica. Siglos XVI-XVIII*), an evolution of the *Collection of maps, plans and drawings* of the General Archive of Simancas (XVI-XVIII centuries), whose objective was to develop tools to support semantic annotation tasks, information search, extraction of hidden relationships in the texts and visualization of the results to facilitate the research of historians. Through these two examples, this article seeks to show the methods, processes, and possibilities of success in complex research problems in the Humanities solved with Natural Language Processing (NLP) techniques and data analysis.

KEYWORDS: Artificial Intelligence, Digital Humanities, Natural Language Processing, Web Applications, Data Analysis.

RHD 7 (2022)

ISSN
2531-1786

10.5944/rhd.vol.7.2022

Revista de Humanidades Digitales: http://revistas.uned.es/index.php/RHD
Recibido 15/09/2021 – Aceptado 30/03/2022

1. INTRODUCCIÓN[1]

Como es sabido, las Humanidades Digitales (HD) son un área de investigación, enseñanza y creación en la que convergen las Humanidades y la Informática y en la que se trabaja sobre materiales previamente digitalizados u originados en el medio digital (McGillivray et al., 2020a; Spence, 2014; Rojas, 2017).

Se considera que las HD se inician con los trabajos de investigación del padre Busa (del Rio Riande, 2014), pero solo es en los comienzos del siglo XXI cuando surge la necesidad de nombrarla como un campo, pues comienzan a necesitarse criterios e indicadores sobre el tipo de conocimientos y habilidades tanto para la descripción de perfiles académicos como para la formación de investigadores y profesionales.

Para delimitar estos saberes es necesario visitar otras disciplinas muy relacionadas con las HD como la inteligencia artificial (IA), el procesamiento del lenguaje natural (PLN) y el análisis de datos o minería de textos (text mining, en inglés), entre otras. La formación en HD debe capacitar para actividades en las que el humanista digital sea capaz de hacer compatibles las necesidades de soporte informático de los humanistas con los recursos y herramientas informáticas existentes o susceptibles de diseño e implementación: la anotación de corpus literarios y textuales, la catalogación de documentos para repositorios virtuales, la creación de contenidos digitales, o la gestión de bibliotecas físicas y virtuales.

Las HD son interdisciplinares y exigen la convivencia y colaboración entre los profesionales e investigadores de diferentes áreas de conocimiento, pero sobre todo de humanistas e informáticos. Para que esta colaboración no se traduzca en insatisfacción personal, por la insuficiencia de los desarrollos y resultados o el incumplimiento de las expectativas, es necesario ser consciente de algunas cuestiones importantes. Por ejemplo, en un proyecto lo primero es analizar si la tarea o problema planteado por el humanista tiene una solución con los recursos y herramientas tecnológicas disponibles, ya sean de software libre, de código abierto o propietario. No se debe considerar, a priori, que se puedan resolver todas las tareas o retos planteados por los humanistas con soluciones informáticas existentes. Conocer el problema e identificar la tecnología necesaria exige la comunicación efectiva y procesos colaborativos en los que compartir conocimientos especializados, terminología, y, sobre todo, entender las expectativas de las personas de la otra disciplina. Se pueden reutilizar recursos, técnicas y herramientas, pero posiblemente también haya que desarrollar software específico para los intereses de la investigación o el ejercicio de la profesión en las HD.

Este artículo comienza contextualizando las HD en el marco de la IA, el PLN y el análisis de datos (apartado 2), las tres áreas de la Informática que más aparecen en proyectos actuales de Humanidades. A continuación, se muestra cómo los avances tecnológicos y las

[1] Una parte de este trabajo se presentó por la primera autora en la conferencia online del 23 de abril de 2021 en el Centro de Inteligencia Artificial de Orense en su segundo ciclo de conferencias *Los viernes con IA*. Accesible desde: https://www.youtube.com/watch?v=pD3pFhlrvXQ&t=346s.

técnicas de IA en general se incorporan en proyectos de Humanidades de la mano de un caso de éxito de dos investigadores (apartado 3) y a través de un caso de estudio práctico de implantación de diferentes tecnologías en el corpus digital DIMH (El Dibujante Ingeniero al servicio de la Monarquía Hispánica. Siglos XVI-XVIII) (apartado 4). En el primero de los casos la tarea a resolver era de atribución de la autoría de las novelas de la novelista italiana con seudónimo Elena Ferrante, y en el segundo el principal objetivo del proyecto era encontrar el soporte informático que facilitara la investigación de los historiadores del arte. En ambos casos es necesario procesar información textual, gestionar y estructurar datos e información con tecnología adecuada que permita la visualización o la comprensión del contenido del corpus mostrando aspectos ocultos para las metodologías tradicionales que faciliten el trabajo de los humanistas.

2. INTELIGENCIA ARTIFICIAL, PROCESAMIENTO DEL LENGUAJE NATURAL Y ANÁLISIS DE DATOS

Hay una definición de IA que explica que esta disciplina Informática sea considerada de gran importancia y que, a su vez, genere muchas expectativas que no siempre se hacen realidad. Esta definición entiende que las máquinas adquieren inteligencia a través de los algoritmos, de forma que se dice que un agente o sistema o pieza de software de IA realiza actividades humanas como la percepción de su entorno, el razonamiento con información, el aprendizaje y la resolución de problemas (Inteligencia artificial, 2021). Pero hay que recordar que estas actividades humanas las realiza una máquina de forma muy similar a como un avión vuela con respecto un ave. Quizá sea más acertado decir que un sistema es de IA, cuando muestra una capacidad de interpretar datos externos, lo hace correctamente o aprende automáticamente de ellos y los emplea para realizar tareas concretas. Una vez que estas labores se definen y se concretan algunas aproximaciones que las resuelven, surgen nuevas áreas de la Informática. Es lo que ocurrió con el procesamiento de palabras y frases en textos o conversaciones y con el reconocimiento de caracteres en imágenes, que se convirtieron en dos áreas de la Informática, PLN y el reconocimiento de imágenes, respectivamente. Así, en el área de la IA suelen quedar todos aquellos problemas o tareas de los que se desconoce su solución automática.

Como ejemplos del cumplimiento de expectativas generadas por los avances en la investigación y desarrollo en Informática y en concreto en la IA, a continuación, se organizan cronológicamente aplicaciones desarrolladas con diferentes tecnologías, para cuestionar su potencial utilización en las HD:

- Década de 1980: Se popularizan los ordenadores personales y con ellos la ofimática, la aplicación de las bases de datos para compañías y los programas de cálculo específicos, como los de cálculo de estructuras para arquitectos e ingenieros.
- Década de 1990: Se generalizan la información en la web, los recomendadores de películas o productos a partir de nuestros datos y comportamientos como usuarios en

internet. Se desarrollan aplicaciones comerciales complejas como son los sistemas informáticos de diseño para fabricantes de moda, robotización de las fábricas y almacenes de stock de productos, la recolección de pistachos con robots o la gestión de inversiones.

- Primera década del siglo XXI: Se produce el *boom* de las redes sociales, el consumo de las noticias cambia, se disparan las compras online mediante tiendas virtuales como Amazon, se usan habitualmente los servicios de la banca online.
- Segunda década del siglo XXI: El traductor de Google acapara parte del mercado de traducción, se populariza la televisión de pago (Netflix y otras plataformas), el móvil inteligente desbanca a otros dispositivos como los relojes de muñeca, la criptomoneda aumenta su presencia en los mercados.

Si trasladáramos esta cronología a las HD, podríamos considerar que nos encontramos a finales de los noventa, aunque seguramente en breve veremos cómo la IA permitirá desarrollar aplicaciones que actualmente ni nos imaginamos (Herranz et al., 2019).

En el informe El Estado de Inteligencia Artificial 2020, elaborado por McKinsey & Company (Portaltic, 2021), a partir de una encuesta a 2.395 industrias y empresas de todo el mundo, se muestra que la adopción de técnicas de IA aumentó, aunque no masivamente, en 2020 y que en el 22% de las empresas que ya habían implementado algún tipo de IA se observó un impacto positivo sobre su EBIT (beneficio antes de intereses e impuestos). En el informe de Microsoft (2019), elaborado a partir de información de 277 empresas en 7 sectores y 15 países de Europa, se concluye que el 71% de las empresas europeas considera que la IA es un tema importante para los directivos y un 57% de las compañías espera que la IA tenga un alto o muy alto impacto en las áreas de negocios que son totalmente desconocidos para las compañías en la actualidad. Aunque solo el 4% de las empresas está utilizando activamente la IA en procesos y tareas llamadas avanzadas, el 28% de las empresas encuestadas se encuentran en una etapa en la que usan la IA en procesos en la empresa. Finalmente, en el informe de Microsoft España (2020), IA en el sector público: Perspectivas europeas para 2020 y años siguientes (España), se dice que el 33% del sector público en España ya ha implementado alguna solución de IA y que, aunque su uso en el ámbito público se encuentra en una fase inicial en nuestro país, el 80% reconoce que la implantación de esta tecnología es una prioridad digital. No nos atrevemos a sugerir cuál será la implantación de las HD en la empresa en los próximos años, pero es una realidad que en la investigación ya está implantada. Necesitaremos unos años para observar la implantación real de la IA en la tercera década del siglo XXI y el cumplimiento de las expectativas.

En PLN, el área del procesamiento automático del lenguaje natural que pretende simular el conocimiento y el comportamiento humano en el dominio del lenguaje y la comprensión lingüística, se desarrollan recursos de información y algoritmos para resolver problemas lingüísticos (Piotrowski, 2012), pero también relacionados con la comunicación

inteligente. En este sentido, hay ejemplos bien conocidos, como los diccionarios digitales, los asistentes de voz, como Alexa o Cortana, traductores como el de Google o analizadores sintácticos y morfológicos integrados en buscadores como Google, Bing (Microsoft) o Safari (Apple).

De acuerdo con el informe de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN, 2020), la simulación de procesos cognitivos humanos en una máquina sigue siendo imposible porque sigue sin resolverse el problema científico de comprender cómo funciona el intelecto humano. Sin embargo, informes de varias consultoras pronostican un gran crecimiento del mercado mundial, basándose en la explosión de aplicaciones observada en los últimos años y en un crecimiento exponencial de los datos digitales no estructurados. Así, el Informe afirma:

> Desarrollar las Tecnologías del Lenguaje (TL) para el español es clave: el español es la segunda lengua con más hablantes nativos en el mundo y la tercera por número de hablantes, posición que se mantiene en los rankings de lenguas más usadas en las redes sociales (SEPLN, 2020, p. 5).

Otra de las áreas de la Informática y de la IA en general que aparece en las HD es el análisis de los datos con técnicas estadísticas básicas o con aproximaciones más avanzadas que exigen la disponibilidad de corpus digitales de gran tamaño y convenientemente anotados para facilitar a los algoritmos ejemplos de las regularidades e irregularidades del corpus (García-Serrano & Menta, 2021). Con el big data (ver uso en HD, por ejemplo, en Espino, 2020) para el análisis masivo de datos y el aprendizaje profundo, tenemos que responder a nuevas preguntas, ¿cuándo dispone y maneja el humanista un corpus digital de un tamaño enorme (de volumen equivalente a todas las páginas web del dominio .es, por ejemplo) para aplicar estas nuevas tecnologías? ¿Será suficiente trabajar con un corpus genérico para aplicar los resultados a problemas con corpus concretos? Cuando conozcamos estas respuestas será el momento de plantear el software necesario para aplicarlo también a las humanidades.

A continuación, se muestran dos casos de estudio con un corpus de tamaño medio, casos en los que se aplican técnicas que provienen de las áreas de la Informática citadas anteriormente: la IA, el PLN o el análisis de datos.

3. INVESTIGACIÓN DE ÉXITO: UN AUTOR MISTERIOSO ENTRE NOVELISTAS ITALIANOS

En este apartado se presenta una investigación o caso de estudio que acabó en éxito usando tecnología para resolver un problema de HD con técnicas de PLN y análisis de datos. El problema que se pretendía resolver era una atribución de autoría para una serie de novelas italianas de gran éxito de ventas y que se publicaban bajo pseudónimo. Para desvelar este misterio Tuzzi y Cortelazzo (2018) explotan técnicas y herramientas de estilometría y de agrupamiento (clustering) y clasificación para encontrar evidencias y al autor de un grupo de novelas o atribución de autoría.

Una metodología científica prevé la aplicación de diversos enfoques, métodos y acciones encaminados a la obtención de nuevos resultados científicos o a la constatación de las hipótesis y tesis planteadas en un problema dado. En este caso, el problema, como dijimos, es identificar al autor, autora o grupo de autores que publica bajo el seudónimo de Elena Ferrante novelas superventas. Los métodos o técnicas que se utilizan son análisis estilométricos, y análisis de datos, basados en la similitud o distancia medible entre objetos que se agrupan de acuerdo con sus propiedades. El inicio de la investigación es el desarrollo de un corpus digital con 150 obras de 40 autores italianos (modernos), entre los que se destacan 13 mujeres novelistas y 11 autores de una determinada región italiana.

Una fuente muy importante de información para el análisis de corpus textuales (text mining) en el área de la Informática son las referencias PAN (Webis Group, 2021), un foro internacional en el que se comparan y evalúan los resultados obtenidos por diferentes investigadores al aplicar sus algoritmos y aproximaciones al mismo corpus y en una misma tarea bien definida. En PAN el objetivo es realizar análisis forenses de textos digitales utilizando diferentes técnicas, incluidas las basadas en el análisis estilométrico (Kestemont et al., 2020). Algunas de estas tareas son de atribución de autoría (dado un documento y un conjunto de autores candidatos, determinar cuál de ellos es el autor del documento), verificación de autoría (dado un par de documentos, el objetivo es determinar si están escritos o no por el mismo autor), u ofuscación de autoría (dado un documento y un conjunto de documentos escritos por el mismo autor, parafrasear el primero para que no se pueda identificar al autor).

Para el análisis de la autoría se estudia el estilo de la escritura del autor, porque las personas dejan un patrón característico de cómo expresan sus pensamientos a través del lenguaje escrito. Estos patrones tienen una huella cognitiva que se puede detectar gracias al estudio de las características estilísticas (Pokhriyal et al., 2017). Pueden ser características léxicas y de caracteres, características sintácticas y semánticas que requieren un análisis lingüístico más profundo, o bien las características específicas de cada dominio de estudio o del lenguaje utilizado.

En el misterio de los novelistas italianos, los autores utilizan la teoría léxica de Correspondence Analysis (CA) (Murtagh, 2017) que agrupa objetos (palabras, párrafos o novelas) de acuerdo con su similitud. Se construyen diversas tablas de contingencia, o matrices de frecuencias. Por ejemplo, una matriz de frecuencias contiene para cada novela la frecuencia de aparición de cada palabra a considerar. El estudio también construye las tablas de contingencia con la frecuencia de cada una de las palabras que aparecen en todas las novelas de cada uno de los cuarenta autores. Las novelas consideradas para cada autor forman cuarenta sub-corpus representativos de cada autor.

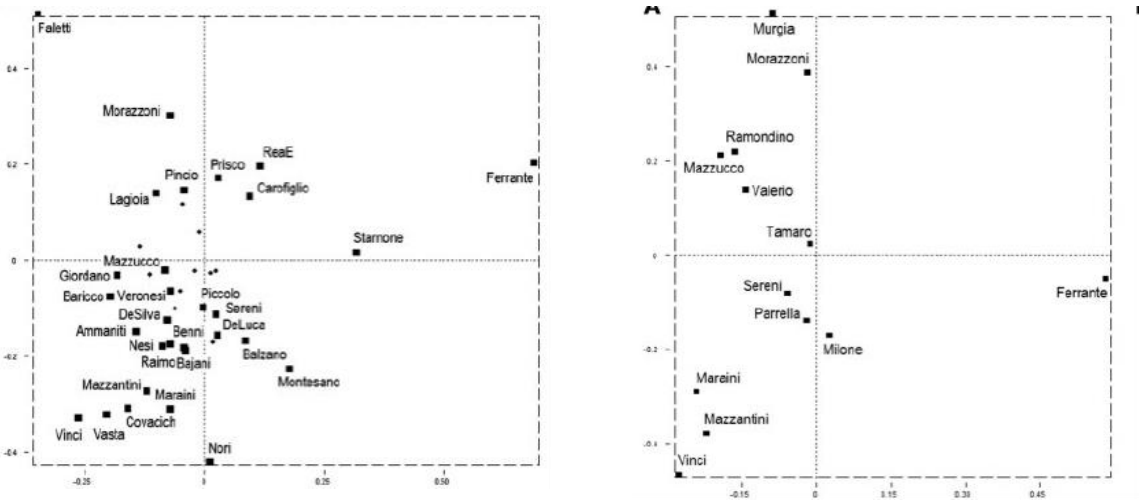
El análisis CA facilita la visibilidad de los resultados al transformar en coordenadas de un espacio multidimensional cada una de las tablas de contingencia construidas, convirtiendo la

distancia euclídea en una distancia chi-cuadrada[2] . Así se descubren relaciones ocultas en los textos del corpus, e incluso nuevo conocimiento, una vez que se ha planteado el problema en términos de la teoría CA.

Aunque se dejan los detalles a los autores Tuzzi y Cortelazzo (2018), a continuación, se indican las pistas principales con las que científicamente concluyen sobre el problema de atribución de autoría a la novelista Elena Ferrante.

El corpus de estudio lo forman casi 10 millones de tokens[3], que provienen de 150 obras, publicadas entre 1987 y 2016 por 40 novelistas italianos. Elena Ferrante ha publicado siete novelas (en 1992, 2002, 2006, 2011, 2012, 2013 y 2014), siendo las cuatro últimas una tetralogía que cuenta la historia de dos mujeres napolitanas. En el corpus hay 13 mujeres novelistas (con 47 novelas) y 11 autores de una determinada región italiana (Campania) que aportan 39 novelas.

La aplicación de la técnica CA produjo una tabla con 24.846 tipos de palabras con una frecuencia mayor que 20. También construyeron 40 tablas de palabras y novelas para cada autor. Con todo ello, los análisis mostraron que Ferrante era una novelista ciertamente peculiar.



(a) (b)

Figura 1. Visualizaciones de CA (Tuzzi & Cortelazzo, 2018). Se muestra en (a) los resultados para los 40 autores (30 puntos con nombre y 10 puntos pequeños sin él) y en (b) los resultados para las 13 novelistas.

En la figura 1(a) se muestra que Elena Ferrante fue una novelista diferente al resto, quedando en la gráfica solo relativamente cerca de Starnone. A continuación, se analizó con CA el caso para las 13 mujeres y las 50 novelas. En la figura 1(b) se observa de nuevo que

[2] Es un aspecto estadístico de las distribuciones de datos, de forma que a partir de las frecuencias de las celdas de la tabla se puede observar una mayor distancia chi-cuadrada en una celda que evidencia una asociación entre las categorías de fila y columna más fuerte de lo que se esperaría.

[3] Escriben, en este sentido, Tuzzi y Cortelazzo (2018): "The novels' dimensions in terms of word tokens, word types, and, as a rough measure of lexical richness, an estimation of the mean number of word types per 1,000-word tokens (based on repeated measures of the number of word types in samples of 1,000 text chunks of 1,000-word tokens in length) show that the novels approximately resemble each other" (p. 687).

Ferrante está a mucha distancia de cualquiera de las otras novelistas. Finalmente analizaron el caso de los 11 autores de la región italiana de Campania, obteniendo similares conclusiones: Elena Ferrante solo está cerca de Starnone.

Pero estos resultados no eran concluyentes para atribuir la autoría a ninguno de los novelistas estudiados. Tenían que seleccionar alguno de los diferentes métodos de atribución de autoría existentes, que dependen del tipo de texto a analizar, la calidad del corpus y los objetivos del análisis. Seleccionaron la técnica estadística de Hierarchical Aglomerative Clustering (HAC) (Kunenets, 2016) con el fin de dividir un conjunto de objetos en grupos homogéneos utilizando un principio de similitud[4].

En el estudio, los objetos son 150 series de palabras de igual longitud (chunks) y tal y como aparecen en la novela (10.000 es el número de palabras del chunk en el estudio). El proceso de selección de chunks y de su comparación de dos en dos se realiza 500 veces y con estas repeticiones se obtienen las distancias entre las novelas. Estas distancias son el criterio principal para disponer las novelas en una clase o grupo en particular. Estas clasificaciones son formas simples y novedosas de entender la estructura de los objetos para identificar los raros porque no pertenecen a estos grupos, o de permitir al agrupar disminuir el volumen de información, por ejemplo, identificando las características más representativas de los grupos y ocultando el resto[5].

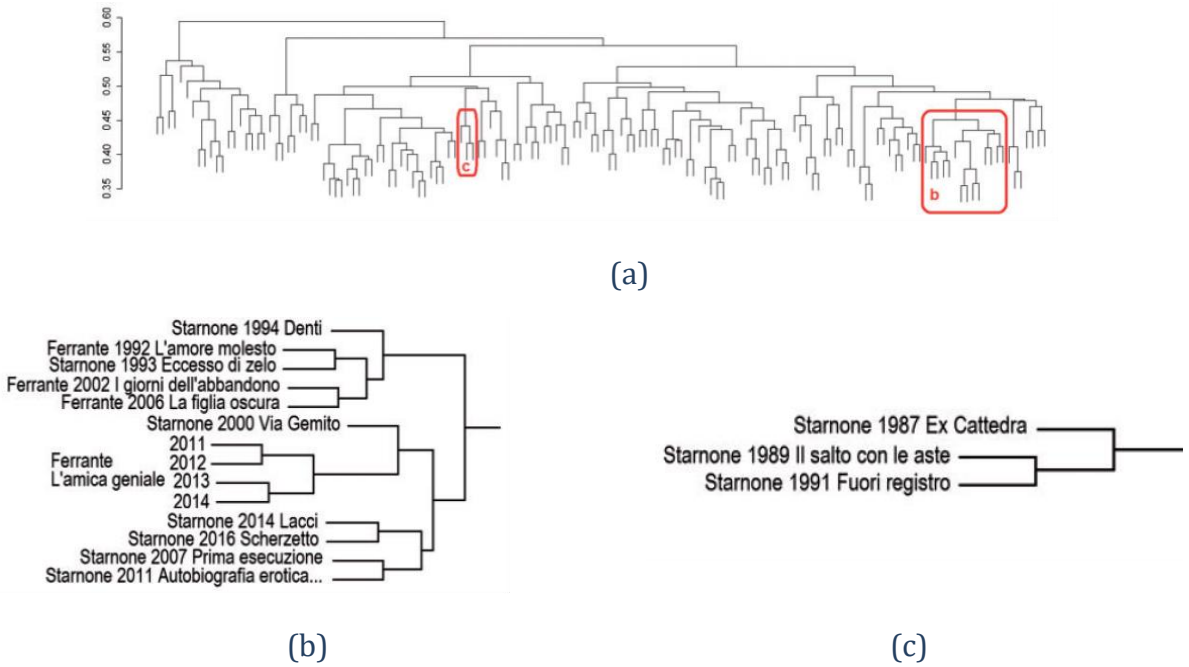


Figura 2. Visualizaciones de HCA (Tuzzi & Cortelazzo, 2018). Se muestra en (a) el dendograma obtenido; en (b) la parte señalada en (a) como b y respectivamente en (c).

[4] Es un tipo de clasificación multidimensional en la que los objetos que deben ser clasificados están definidos por un vector (conjunto) de características o propiedades individuales con los que se construye una tabla objeto y propiedad, sobre la base de la que se construye la matriz o tabla de distancias (similitudes).

[5] Esta técnica se aplica cuando hay mucha cantidad de objetos descritos por las mismas características.

Pero, además, este método permite mostrar el resultado con un tipo de gráfico especial (figura 2) denominado dendrograma (representación en forma de árbol donde cada nivel indica una subcategoría del anterior) que muestra la cercanía o similitud entre los objetos que se clasifican. En estos gráficos las distancias entre los objetos de una agrupación siempre son menores que la distancia mínima entre objetos de dos agrupaciones diferentes. En este caso, las hojas del dendograma son las 150 novelas.

En la figura 2a se distinguen dos grupos, el que incluye a todas las novelas de Ferrante y las de Starnone a partir de 1992 (figura 2b), pues sus tres primeras están en un grupo separado (figura 2c). Pero, a pesar del dendograma, todavía había aspectos que no se podían descubrir con esta representación. Por ello, los autores pasaron a una nueva fase de su investigación: el estudio de matrices de distancia con la perspectiva de un sistema de ordenación (ranking). Para el ranking se utilizaron las distancias entre cada par novelas de la matriz cuadrada de distancias (con 150 filas y columnas) construida iterativamente (como se ha indicado anteriormente con 500 repeticiones). Como las distancias de las filas (o columnas) de la matriz indican la distancia de una novela al resto de novelas, dada una novela, el resto de las novelas puede ser ordenado en relación con la escogida desde la más cercana (distancia menor) a las siguientes.

Este tipo de ordenación puede verse en la tabla de la figura 3, donde se ordenan las 20 novelas más cercanas a las de Elena Ferrante, de acuerdo con la distancia intertextual calculada iterativamente sobre las ocurrencias de palabras gramaticales (artículos, preposiciones, conjunciones y pronombres). De esta forma pretendían eliminar la influencia de las palabras con contenido y, con ello, la similitud entre las novelas de Ferrante y Starnone pasa a ser evidente.

L'amore molesto 1992	I giorni dell'abbandono 2002	La figlia oscura 2006	L'amica geniale 2011	Storia del nuovo cognome 2012	Storia di chi fugge e di chi resta 2013	Storia della bambina perduta 2014
Starnone 1993	Ferrante 2006	Ferrante 2002	Ferrante 2012	Ferrante 2011	Ferrante 2014	Ferrante 2013
Ferrante 2006	Starnone 1993	Starnone 1993	Ferrante 2014	Ferrante 2014	Ferrante 2012	Ferrante 2012
Ferrante 2002	Ferrante 1992	Ferrante 2013	Ferrante 2013	Ferrante 2013	Ferrante 2011	Ferrante 2011
Starnone 1994	Starnone 2016	Ferrante 2014	Ferrante 2006	Ferrante 2006	Ferrante 2006	Ferrante 2006
Ferrante 2011	Starnone 1994	Ferrante 2012	Ferrante 1992	Starnone 1993	Starnone 2014	Starnone 2014
Ferrante 2012	Ferrante 2013	Ferrante 1992	Starnone 1993	Starnone 2014	Starnone 1993	Starnone 1993
Starnone 2000	Starnone 2007	Ferrante 2011	Starnone 2000	Starnone 2011	Starnone 2016	Starnone 2016
Ferrante 2014	Ferrante 2012	Starnone 1994	Starnone 2014	Starnone 2016	Starnone 2011	Ferrante 2002
Ferrante 2013	Ferrante 2014	Milone 2015	Carofiglio 2004	Starnone 2000	Ferrante 2002	Starnone 2011
Milone 2015	Murgia 2015	Starnone 2011	Milone 2015	Carofiglio 2011	Murgia 2015	Ferrante 1992
Carofiglio 2004	Ferrante 2011	Starnone 2007	Starnone 2011	Ferrante 1992	Carofiglio 2011	Tamaro 1994
De Luca 1998	Milone 2015	Tamaro 1994	Balzano 2014	Ferrante 2002	Starnone 2000	Murgia 2015
Murgia 2015	Tamaro 1994	Starnone 2014	De Luca 1998	Milone 2015	Carofiglio 2010	Starnone 2000
Starnone 2007	Starnone 2011	Starnone 2016	Ferrante 2002	De Silva 1999	Starnone 2007	Piccolo 2008
Starnone 2016	De Luca 1998	De Luca 1998	Sereni 2015	Murgia 2015	Milone 2015	Carofiglio 2011
Starnone 2011	Starnone 2014	Murgia 2015	Carofiglio 2011	Balzano 2014	Ferrante 1992	Carofiglio 2004
Starnone 2014	Starnone 2000	Carofiglio 2004	Starnone 2016	Carofiglio 2004	Tamaro 1994	Carofiglio 2010
Milone 2013	Carofiglio 2004	Balzano 2014	Murgia 2015	De Luca 1998	Balzano 2014	Carofiglio 2003
Tamaro 1994	Balzano 2014	Tamaro 1991	Piccolo 1996	Piccolo 2008	Carofiglio 2004	Milone 2015
Carofiglio 2003	De Luca 2011	De Luca 2011	Carofiglio 2003	Carofiglio 2010	Carofiglio 2006	Carofiglio 2006

Figura 3. Muestra de la proximidad entre obras del corpus (Tuzzi & Cortelazzo, 2018).

Además de los métodos cuantitativos anteriores, también realizaron un estudio cualitativo con el que descubrieron que había cuatro palabras nada frecuentes en italiano, pero sí en las novelas de Ferrante y Starnone, precisamente unas palabras de la región italiana de Campania (de donde es Starnone) y otras utilizadas con diferentes variantes y concluyeron que:

> Investigando únicamente con las palabras gramaticales, eliminando así las posibles similitudes debidas a factores ligados a los contextos históricos, sociales y culturales y al contenido general de las novelas, se refuerza la similitud entre Starnone y Ferrante, como si se tratara de la misma mano… es bastante difícil imaginar que Starnone no haya jugado ningún papel en la planificación y / o redacción del trabajo de Ferrante.
> Es difícil definir con precisión su papel: también podría ser solo una de las manos y cabezas que han conformado el fenómeno Ferrante, pero de alguna manera ha dejado su huella en él (Tuzzi & Cortelazzo, 2018, pp. 697-698).

Los dos filólogos investigadores opinan que estas pistas obtenidas no son del todo concluyentes, pero sí muy indicativas de la cercanía entre la novelista superventas con pseudónimo Elena Ferrante y el novelista italiano Domenico Starnone. Pero además comentan que se sabe que los derechos de autor de las novelas de Elena Ferrante los cobra la pareja real de Starnone, que también es escritora, aunque hace años que no publica nada.

4. LA INVESTIGACIÓN EN EL CORPUS DIMH

En el contexto DIMH El dibujante ingeniero al servicio de la monarquía hispánica. Siglos XVI-XVIII (HAR2012-31117)[6], inicialmente financiado por el Ministerio de Economía y Competitividad entre 2013 y 2015 y en cuyo marco continuó la investigación durante mucho más tiempo, se presentan en este apartado los resultados obtenidos en relación con el soporte informático a la investigación en historia del arte (García-Serrano & Castellanos González, 2016). En este caso de estudio se describen brevemente las diferentes etapas en las que se ha trabajado sobre el corpus digital DIMH y algunas de las herramientas diseñadas para soporte a la investigación humanista.

La construcción de corpus digitales específicos para las HD es un problema en el que se ha trabajado ampliamente (Martínez Cantón, 2017; Ueda et al., 2020). El corpus digital con el que se comienza a trabajar en el proyecto DIMH ya estaba desarrollado: es la Colección de mapas, planos y dibujos[7] del Archivo General de Simancas (AGS), que contiene 7.792 fichas textuales digitales relacionadas con mapas, planos y dibujos (en formato de imagen), a las que es posible acceder con búsquedas simples o avanzadas según los campos descriptivos (metadatos) de las fichas[8]. La información de las fichas originales está semiestructurada, ya que, por una parte, está codificada con metadatos en el formato estándar RDF-DC y, por otra, hay metadatos con información en texto libre donde se encuentra una parte importante de la

[6] Proyecto de Investigación Fundamental No Orientada (VI Plan Nacional de I+D+i 2008-11, Convocatoria 2012). Investigadora Principal: Alicia Cámara Muñoz (UNED). Accesible desde: http://dimh.hypotheses.org/.
[7] Accesible desde: http://www.mcu.es/ccbae/es/mapas/principal.cmd.
[8] Accesible desde: http://www.mcu.es~mapas_planos_dibujos&posicion=1&id=30485#.

información de interés para los historiadores, que son los usuarios potenciales del sistema automático.

Como sabemos, los metadatos son datos que describen a los datos y hacen que las descripciones de los contenidos sean uniformes. Un ejemplo es el siguiente extracto de una ficha, donde todos los términos entre "<" y ">" son del metalenguaje que describe los metadatos:

```
<dc:title>Traza del castillo que se ha de hacer en la Aljafería de Zaragoza, por Tiburcio Spanochi</dc:title>
<dc:description> Realizada con motivo de la publicación de los tomos XIX y XXIII de la Historia de España dirigida por Ramón Menéndez Pidal</dc:description>
<dc:type>Vídeos y Diapositivas</dc:type>
<dc:language>spa</dc:language>
<dc:date>[enero 1989]</dc:date>
<dc:creator>Espasa Calpe, S.A.</dc:creator>
<dc:identifier>http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=3405</dc:identifier>
<dc:rights>The Public Domain Mark (PDM)</dc:rights>
```

Una de las primeras actuaciones que se realizan en los corpus digitales en las HD es identificar las palabras de contenido denominadas entidades nombradas, que se refieren a nombres propios, localidades, organizaciones y otras para indicar su aparición mediante anotaciones. Este proceso de anotación enriquece el contenido textual de las fichas, en este caso con (a) las entidades nombradas (codificado por <nes>) y sus categorías (<nes_person>, <nes_organization>, <nes_location>, <nes_misc>), (b) los sintagmas nominales (<sintagmas>) y (c) los lemas (<lemas>). El resultado de realizar el proceso automático de enriquecimiento en cada ficha proveniente de la descarga de las fichas de esta colección en ficheros independientes es un nuevo fichero en formato XML con nuevos metadatos, que constituye un nuevo corpus, denominado DIMH:

```
<Ficha id="179110">
<Materia>España-Aragón-Zaragoza (Provincia)-Zaragoza</Materia>
<Tipo>Mapas</Tipo>
<Referencias>Referencias: Mapas, planos y dibujos (Años 1503-1805). Volumen I : p. 1018</Referencias>
<Creador> Spannocchi, Tiburzio (1541-1606) </Creador>
<Notas>Tinta y color amarillo</Notas>
…
<Notas>AGS. Guerra y Marina, Legajos, 00354. Con carta de Tiburcio Spanochi al Rey, Jaca, 22 de julio de 1592</Notas>
<Publicacion>1592</Publicacion>
<Notas>Manuscrito sobre papel. </Notas>
<Titulo>Traza del castillo que se ha de hacer en la Aljafería de Zaragoza [Material cartográfico]</Titulo>
…
<nes>Zaragoza Zaragoza Madrid Tiburcio_Spanochi Spannocchi MenÉndez_Pidal Historia_de_España Tiburzio Nerea Muñoz AGS Madrid Jaca Felipe_II España Cámara Aragón Alicia Aljafería_de_Zaragoza </nes> …
</Ficha>
```

Una vez enriquecida y representada la información de las fichas originales de la colección de mapas, planos y dibujos de AGS, se desarrolló una aplicación web para búsquedas sobre el contenido, tanto basadas en palabras clave como configurables por el

usuario, sobre algunos de los parámetros de la estructura de metadatos de los documentos del corpus o sobre el tipo de información lingüística contenida en el corpus (García-Serrano et al., 2015). Para desarrollar un buscador hay que realizar un proceso automático que se denomina indexación y una interfaz web para las preguntas del usuario y la presentación de los resultados.

Por otra parte, para la representación del texto libre de las fichas (y poder procesarla posteriormente) se seleccionan las palabras o términos que lo describen y distinguen mejor (evitando palabras demasiado genéricas). Para ello se aplicó el algoritmo de Kullback-Leibler Divergence (KLD), que analiza la distribución de una palabra tanto en un documento como en la colección (Castellanos González et al., 2012).

Con el buscador desarrollado, los historiadores accedían al contenido de las fichas y a las nuevas anotaciones. Para responder a la siguiente pregunta de investigación ¿se puede clasificar o agrupar toda esta información para facilitar la investigación de los historiadores del arte? Se desarrolló una aplicación que utilizando una técnica analítica identificara automáticamente la estructura no visible de la información (sus relaciones). Para ello, hay diferentes aproximaciones para la agrupación de documentos teniendo en cuenta su contenido.

Uno de los enfoques analíticos muy utilizado en HD se basa en el análisis de las características del contenido mediante la aplicación de técnicas probabilísticas, principalmente Latent Dirichlet Allocation (LDA), (Meeks & Weingart, 2012). En la teoría LDA se supone que existe una estructura latente en los datos (definida por los temas tratados) que está relacionada con la distribución de palabras y que permite clasificar los contenidos. En el trabajo de Yang et al. (2011) se detecta el conjunto de temas (topics) abordados en una colección de periódicos históricos. Otros trabajos interesantes en HD en español son los de Cebral Loureda (2020), en un corpus de textos clave de la modernidad filosófica, y el de Vivó Capdevila (2021) para un corpus literario.

Aunque las técnicas probabilísticas han demostrado su potencial, su desventaja más importante está relacionada con la dificultad de visualización e interpretación de la agrupación obtenida. Y la visualización de los resultados obtenidos automáticamente es el aspecto crítico que genera credibilidad de las técnicas y permite el progreso de las herramientas para los usuarios. Para resolver este problema, en el caso de estudio DIMH se propuso una organización de contenido basada en una representación conceptual de los datos, el Análisis formal de Conceptos (FCA, en sus siglas en inglés) para inferir relaciones entre los datos y organizarlos de acuerdo con estas relaciones (Castellanos González et al., 2017a). Otra de las ventajas de FCA sobre LDA es que no es necesario decidir a priori el número de conceptos o nodos de la estructura en red que se genera. FCA (Wille, 1992) genera automáticamente un conjunto de conceptos formales donde un concepto formal incluye un grupo de objetos que comparten el mismo conjunto de características y organiza estos conceptos formales en una representación jerárquica denominada retículo o red (nodos en el espacio con algunas relaciones o uniones de dos en dos por aristas) con propiedades teóricas especiales,

que tiene ventajas en comparación con otras representaciones menos formales y que permiten la aplicación de FCA en escenarios donde los enfoques probabilísticos o basados en gráficos no son adecuados (Schmidt, 2012).

En DIMH se diseñó e implementó el modelo basado en FCA sobre la base de conceptos formales o grupos de objetos, en este caso fichas, siendo sus atributos las palabras elegidas automática o manualmente (términos). El proceso consiste en:

(1) Extracción de información de las fichas: Se extrajo solo el contenido de los metadatos siguientes: publicación, referencia, notas, entidades nombradas <nes>, temática, materia y título. Se aplicó un proceso de eliminación de palabras vacías (sin contenido).

(2) Creación del contexto formal: mediante una matriz de adyacencia (tabla) que indica cuándo una palabra (término o atributo) aparece o no en el contenido de cada ficha (objeto).

(3) Reducción del contexto formal: El contexto formal generado incluye toda la terminología susceptible de representar a las fichas. Sin embargo, esto da lugar a información redundante o poco significativa, por lo que se redujo el contexto formal. La reducción consistió en encontrar aquella terminología que permite identificar más relaciones entre las fichas, sin pérdida significativa de información, mejorando tanto el tiempo de ejecución del algoritmo de FCA como sus resultados. Este refinamiento dio como resultado una organización de los contenidos más específica y, en consecuencia, más informativa.

(4) Ejecución del algoritmo de FCA: Se generan los conceptos formales con el algoritmo de *Next Neighbourhoods* (Cigarrán, 2008). Sin entrar en detalles técnicos, en esta fase se generan todos los grupos posibles (que agrupan fichas que comparten un conjunto de palabras o atributos) y las relaciones de orden entre los mismos.

Realizadas las pruebas de usuario con los historiadores del arte, éstos mostraron su interés en algunas agrupaciones generadas automáticamente por la técnica FCA aplicada al corpus DIMH y echaron en falta otras que ellos deseaban encontrar. Por esta razón en el proyecto DIMH se decidió, por una parte, modificar el modelo FCA para incorporar esta información concreta y, por otra, construir una taxonomía que además de la información contenida en las fichas, incorporara los detalles de interés aportados mediante el conocimiento previo explicitado por los expertos humanistas del proyecto en forma de terminología (Castellanos González & García-Serrano, 2017b).

Para tener en cuenta el conocimiento previo en el modelo FCA, durante la etapa de reducción del contexto formal, cada vez que una de las entidades que aparece en la taxonomía es detectada en el texto de una de las fichas, se obliga a que esta entidad sea incluida como atributo relevante para la ficha, independientemente de su frecuencia de aparición y, por lo tanto, estará en el retículo generado (resultado) tras la aplicación de FCA. Mediante este

refinamiento, toda referencia al conocimiento previo o terminología señalada que aparezca en las fichas será tenida en cuenta y el modelo resultante organizará tanto a la información deseada por los historiadores como al resto de información aportada por las fichas.

(a) (b)

Figura 4. (A) Menú principal de la aplicación FCA y (b) Buscador de los conceptos formales generados automáticamente (FCA). Figura extraída de (García-Serrano & Castellanos González, 2016).

Puesto que para explorar los conceptos formales es necesario desarrollar una interfaz, se decidió una de búsqueda sobre el retículo. Para permitir esta búsqueda, todos los conceptos formales generados se indexaron (tipo de almacenamiento) junto con su información. Por ejemplo, si se quiere buscar información relacionada con mapas, se obtendrían los resultados de la figura 4, donde pueden observarse los conceptos formales asociados a los mapas (con indicación del número de fichas que incorporan). Como se puede ver, se obtiene información agrupada de diferentes ámbitos geográficos de los mapas (Galicia, Andalucía, Madrid, Ceuta), diferentes características (azul, legajo) o diferentes protagonistas (duque, marqués de la Ensenada, magestad, Julián de Arriaga), etc.

En esta segunda etapa del estudio, los historiadores encontraron muy difícil consultar y explotar la información presentada con el buscador, por lo que se procedió al desarrollo de una nueva interfaz con tecnología avanzada en su momento como eran las burbujas de palabras (Filter, 2015) y sobre ella realizar la evaluación de la utilidad de la aplicación con los usuarios finales (figura 5). Tanto la etapa de desarrollo de interfaces como la de evaluación de la aplicación desarrollada con usuarios son cruciales en los proyectos de investigación y también a nivel profesional antes de que una aplicación pase a producción (uso comercial fuera del laboratorio de pruebas). Sin embargo, en los proyectos de investigación no suele haber tiempo ni recursos para realizarlas.

Figura 5. Navegación por conceptos formales basada en nubes de palabras y eventos (Merás et al., 2017).

Tras la evaluación con los usuarios de la aplicación FCA desarrollada en la nueva interfaz, los historiadores se dieron cuenta de que la información y las relaciones que emergían no incluían la dimensión temporal asociada, aspecto importante en su trabajo. En este caso, como en cualquier sistema de acceso a la información, ocurre que, aunque la información temporal perfecta la aportaría un historiador, la contextualización temporal automática de un corpus puede aprovechar recursos LOD (Linked Open Data). Como, además, el grupo de Informática del proyecto DIMH estaba trabajando en temas relacionados con la extracción de información de fuentes en abierto, se propuso una nueva tarea en el mismo caso de estudio: la integración de información temporal en el corpus semiestructurado DIMH, siendo la información temporal extraída automáticamente desde la DBpedia (Bizer et al., 2009; Rettinger et al., 2015).

La propuesta (Merás et al., 2017) es suficientemente general para ser aplicada a diferentes dominios. Consiste en identificar recursos de la DBpedia relevantes al dominio, utilizando las etiquetas clase (rdf:type) y categoría (odcterms:subject) para, a continuación, extraer la información temporal teniendo en cuenta la consistencia entre recursos hermanos (owl:sameAs) en DBpedia (descripción en diferentes idiomas). Con la información temporal extraída de los recursos se alimenta una línea de tiempo y se interseca a su vez con la información temporal extraída del dominio, en este caso el corpus DIMH.

En general, la identificación de fechas en un texto es una tarea relativamente compleja ante la ausencia de uniformidad y normalización de la anotación de la información temporal en cualquier lengua (Vicente-Díez et al., 2010; Vázquez & García-Serrano, 2015), pero, además, en las fichas DIMH hay fechas tanto en español como en inglés, con lo que la extracción de información temporal del corpus conlleva las tarea de tratamiento del multilingüismo (en oraciones cortas), el análisis de colisiones entre las fechas extraídas y la representación de intervalos temporales.

Para anotar automáticamente las fichas del corpus DIMH con los eventos relacionados

en una línea de tiempo (alimentada automáticamente desde la DBpedia) y con la información temporal extraída del corpus se diseña un formato de anotación temporal (.moment), basado en la lógica temporal de intervalos que define operadores para expresar las relaciones temporales entre intervalos (Allen, 1983).

Una vez realizada la anotación temporal del corpus DIMH, se observó que en una ficha se pueden encontrar varios intervalos que se contengan entre sí, además de períodos extremadamente extensos que se deben considerar como ruido. Como el objetivo de la extracción de esta información era anotar y agrupar las fichas en momentos históricos, se diseña e implementa una estrategia que, ante una colisión, selecciona el intervalo más específico. El corpus anotado temporalmente se puso a disposición de la comunidad científica[9]. Un ejemplo parcial de ficha es el siguiente:

```
<Ficha id="183679">
<Tipo>Ilustraciones y Fotos</Tipo>
…
<time_relatedEvents>
<event id="0000000018" title= "Siglo XVIII">
<class> century</class>
<moment start="1701-01-01" end = "1800-12-31" /> </event>

<event id="0000001059" title= "Reinado de Carlos III de España">
<class> monarchs_europe </class>
<class> monarchs_spain </class>
<moment start="1759-08-10" end= "1788-12-14" /> </event>
</time_relatedEvents>
… </Ficha>
```

Para presentar los resultados obtenidos y realizar una búsqueda tanto por la dimensión temporal como por la dimensión léxica, se desarrolló una interfaz con el modelo de visualización de retículos desarrollado previamente. Con la resolución de colisiones, los intervalos temporales suelen ser significativos (comprobación manual) y relacionados con la obra descrita en las fichas (plano, mapa o dibujo). Además, se evita que fichas con intervalos muy grandes creen una ventana temporal muy amplia, provocando una organización de relaciones entre conceptos con atributos temporales muy alejados en el tiempo.

Del trabajo realizado en este caso de estudio se han obtenido resultados como un software relacionado el corpus DIMH y una ontología en Protégé[10] (herramienta open source) desarrollada para una práctica en un curso de máster de investigación en tecnologías de la lengua.

5. CONCLUSIONES

En este trabajo se ha argumentado sobre la influencia de diferentes áreas de la Informática en las HD y se han planteado dos casos de estudio diferentes, uno acaba con éxito, dado que aporta una solución al problema planteado y, en el otro, también exitoso, se

[9] Accesible desde: https://github.com/meras0704/DBpediaTime.
[10] Accesible desde: https://protege.stanford.edu/.

hace un recorrido de la investigación realizada por un equipo de historiadores e informáticos para identificar el tipo de herramientas o aplicaciones informáticas necesarias para facilitar la investigación humanista. En este segundo caso se detallan los objetivos y fundamentación que dirigen la evolución de las herramientas de soporte informático desarrolladas. Fundamentalmente pretendemos mostrar las posibilidades de solucionar problemas complejos con técnicas de PLN y análisis de datos.

En la revisión del estado del arte se han referenciado problemas planteados en el área de las Humanidades y resueltos con una solución basada en alguna técnica del PLN o de análisis de datos, y siempre que se disponga de un corpus digital relativamente reducido para ser computable y suficientemente extenso para poder analizarlo. Este aspecto es un punto de unión de ambos casos de estudio, junto con la variedad de técnicas del procesamiento de lenguaje natural que se puede aplicar para resolver una tarea determinada en el marco de un problema complejo de Humanidades. Características como la volumetría, la tarea a resolver o los recursos externos disponibles son clave para seleccionar un enfoque y muestran que al abordar un proyecto no siempre puede utilizarse la última técnica publicada en investigación (y de actualidad) para resolver cualquier problema.

Ambos casos de estudio demuestran la versatilidad de las técnicas disponibles actualmente, y que van desde la utilización de la estilometría y el posterior agrupamiento de las características obtenidas, hasta la utilización de análisis formal de conceptos y la extracción de entidades nombradas como pueden ser nombres propios o localidades.

En los dos casos se han presentado algunos de los problemas y tareas de investigación que pueden plantearse en las HD, pero, bien es verdad que hay otras tareas específicas que ya forman parte de las HD, como son la digitalización y preservación del patrimonio (Literatura, documentalistas), la geolocalización (Historia y Arqueología), la minería de textos para la detección de noticias falsas (Periodismo), la detección de plagio, la visión artificial (Arqueología o caligrafías en textos antiguos), o bien tareas como la estandarización de sistemas de almacenamiento de los objetos de estudio, la navegación y búsqueda en repositorios específicos textuales o multimedia, el análisis léxico y semántico de textos, las aplicaciones con sistemas GIS, la catalogación web con buscadores facetados o en texto libre y otras aplicaciones de la web semántica.

Sin embargo, como se indica en McGillivray et al. (2020b), hay ciertos aspectos del campo del PLN o de la IA que pueden no ser directamente aplicables en las HD, debido a que puede haber desajustes, por ejemplo, entre la evaluación realizada en PLN y las necesidades de los estudiosos del DH en términos de evaluación y rendimiento de las herramientas. Es el caso del análisis de datos masivos y del deep learning, campos en los que no tienen los mismos objetivos en Informática y en Humanidades. Las dos son aproximaciones de tipo caja negra en la que la casi total ausencia de explicaciones y visualización de los resultados dificulta el análisis final cualitativo de los usuarios, aspecto muy relevante en los proyectos de HD (Sanz Cabrerizo, 2021).

Es necesario seguir haciendo estudios y proyectos en los que se planteen aspectos teórico-prácticos para identificar, entre todas las posibles aproximaciones, técnicas y metodologías útiles para que los humanistas puedan abordar una tarea profesional o de investigación, y para conocer qué tipo de aplicaciones necesitan y se pueden diseñar e implantar en un futuro próximo[11].

REFERENCIAS BIBLIOGRÁFICAS

Allen, J. F. (1983). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, *26*(11), 832-843. https://doi.org/10.1145/182.358434

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, *7*(3), 154-165. https://doi.org/10.1016/j.websem.2009.07.002.

Castellanos González, A., Cigarrán, J., & García-Serrano, A. (2012). Using IR Techniques for Topic-based Sentiment Analysis through Divergence Models. *Workshop on Sentiment Analysis at SEPLN*.

Castellanos González, A., Cigarrán, J., & García-Serrano, A. (2017a). Formal Concept Analysis for Topic Detection: A Clustering Quality Experimental Analysis. *Information Systems*. *66*,24-42. https://doi.org/10.1016/j.is.2017.01.008.

Castellanos González, Á., & García Serrano, A. (2017b). Representación y organización de documentos digitales: Detalles y práctica sobre la ontología DIMH. *Revista de Humanidades Digitales*, *1*, 314- 344. https://doi.org/10.5944/rhd.vol.1.2017.17155.

Cebral Loureda, M. (2020). Voluntad y deseo en la filosofía moderna: Un acercamiento computacional. *Revista de Humanidades Digitales*, *5*, 42-65. https://doi.org/10.5944/rhd.vol.5.2020.27495

Cigarrán, J. M. (2008). *Agrupación de resultados de búsqueda mediante análisis formal de conceptos*. [Ph.D. thesis]. UNED. http://e-spacio.uned.es/fez/view/tesisuned:lnglnf-Jcigarran.

del Rio Riande, G. (2014). ¿De qué hablamos cuando hablamos de Humanidades Digitales?. En: *Humanidades Digitales: Culturas, Tecnologías, Saberes*. Facultad de Filosofía y Letras de la Universidad de Buenos Aires. https://www.aacademica.org/gimena.delrio.riande/90.pdf

[11] Agradecemos la financiación del Ministerio de Ciencia e Innovación CONVOCATORIA 2020, del proyecto CLARA-HD (PID2020-116001RB-C32).

Espino, F. (2020). Big data, criticometría y el estudio de las literaturas nacionales en la bibliografía crítica: El caso excepcional de la literatura cubana. *Revista de Humanidades Digitales*, *5*, 66-85. https://doi.org/10.5944/rhd.vol.5.2020.27625

Filter, J. (2015). *Interactive Visualization of Large Concept Lattices*. [Bachelor Thesis]. Supervisors: Nürnberger, A., & García-Serrano, A. Data and Knowledge Engineering Group. Faculty of Computer Science. OVG Univ. of Magdeburg.

García-Serrano, A., Castellanos González, A., & Cigarrán, J. (2015). De la preservación digital al acceso semántico de documentos históricos. En *II Congreso de Humanidades Digitales Hispánicas - Innovación, globalización e impacto*. Madrid. https://humanidadesdigitaleshispanicas.es/wp-content/uploads/2020/02/Humanidades-Digitales-Hisp%C3%A1nicas.-Innovaci%C3%B3n-Globalizaci%C3%B3n-e-Impacto.pdf

García Serrano, A. M., & Castellanos González, Á. (2016). Conceptualización, acceso y visibilidad de la información en el proyecto DIMH. En A. Cámara Muñoz, *El dibujante ingeniero al servicio de la monarquía hispánica: Siglos XVI-XVIII* (pp. 379-400). Fundación Juanelo Turriano. http://www.juaneloturriano.com/coleccion-digital/lecciones-juanelo-turriano-de-historia-de-la-ingenieria

García-Serrano, A., & Menta Garuz, A. (2021). Orientaciones y evaluación de técnicas en Humanidades Digitales: de la estadística al *deep-learning*. *HDH 2021-Scire Vías. Humanidades Digitales y Conocimiento*. 4-8 octubre. España.

Grabmeier, J., & Rudolph, A. (2004). Techniques of Cluster Algorithms in Data Mining. *Data Mining and Knowledge Discovery*, *6*, 303-360.

Herranz A., Benjamins, R., Torrubia, A., & Larrañaga, P. (2019). De qué serán capaces la inteligencia artificial y el machine learning en 10 años: los mayores expertos nos responden. *Xataka*.

Inteligencia artificial. (15 de noviembre de 2021). En *Wikipedia*. https://es.wikipedia.org/w/index.php?title=Inteligencia_artificial&oldid=139728139

Kestemont, M., Manjavacas, E., Markov, L., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., & Stein B. (2020). Overview of the Cross-Domain Authorship Verification Task at PAN 2020. *Working notes CLEF 2020* - Conf. Labs Eval. Forum, pp. 22-25. http://ceur-ws.org/Vol-2696

Kunenets, N. (2016). The Technology of Hierarchical Agglomerative Cluster Analysis in Library Research. *Econtechmod. An International Quarterly Journal, 5*(1), 35-41.

Martínez Cantón, C. I. (2017). *Poetriae* y el *Arte de la poesía castellana*: bases para la creación de una colección digital de tratados poéticos castellanos. *Magnificat: cultura i literatura medievals*, 4, 21-42.

McGillivray, B., Beatrice, A., Ames, S., Armstrong, G., Beavan, D., & Ciula, A. (2020a). The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from the Alan Turing Institute. *Figshare*. https://doi.org/10.6084/m9.figshare.12732164.v5

McGillivray, B., Poibeau, T., & Ruiz, P. (2020b). Digital Humanities and Natural Language Processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly* 14, 2. https://hal.archives-ouvertes.fr/hal-02970302

Meeks, E., & Weingart S. (2012). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities*, *2*(1), 1-6.

Merás, A., García-Serrano, A., & Castellanos, A. (2017). Extracción de información temporal de la DBpedia: propuesta de integración en un corpus semiestructurado. *Procesamiento del Lenguaje Natural*, *58*, 117-124.

Microsoft España (2019). Inteligencia artificial en España: *Cómo 277 organizaciones se benefician de la IA*. https://info.microsoft.com/WE-DIGTRNS-CNTNT-FY19-09Sep-27-ArtificialIntelligenceinSpain-MGC0003165_01Registration-ForminBody.html?wt.mc_id=AID732606_QSG_BLOG_278541

Microsoft España (2020). *IA en el sector público: Perspectivas europeas para 2020 y años siguientes (España)*. https://info.microsoft.com/rs/157-GQE-382/images/ES-CNTNT-eBook-SRGCM3981-v2.pdf

Murtagh F. (2017). Big Data Scaling through Metric Mapping: Exploiting the Remarkable Simplicity of Very High Dimensional Spaces Using Correspondence Analysis. En F. Palumbo, A. Montanari & M. Vichi (Eds.), *Data Science. Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 295–306). Springer. https://doi.org/10.1007/978-3-319-55723-6_23

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers. https://doi.org/10.2200/S00436ED1V01Y201207HLT017

Pokhriyal, N., Tayal, K., Nwogu, I., & Govindaraju, V. (2017). Cognitive-Biometric Recognition from Language Usage: A Feasibility Study. *IEEE Transactions on Information Forensics and Security*, *12*(1), 134-143.

Portaltic (14 de enero de 2021) La adopción de Inteligencia Artificial no aumentó masivamente en 2020, según un informe de McKinsey. https://www.europapress.es/portaltic/empresas/noticia-adopcion-inteligencia-artificial-no-aumento-masivamente-2020-informe-mckinsey-20210114145739.html

Rettinger, A., Zhang, L., Tran, T., & Chen, W. (2015). Time-Aware Entity Search in DBpedia. *The Semantic Web: ESWC 2015 Satellite Events*.

Rojas Castro, A. (2017). La edición crítica digital y la codificación TEI. Preliminares para una nueva edición de las Soledades de Luis de Góngora. *Revista De Humanidades Digitales*, *1*, 4-19. https://doi.org/10.5944/rhd.vol.1.2017.16379

Sanz Cabrerizo, A. (2021). Para unas lecturas remediadas: análisis cuantitativo y cualitativo de textos. *Revista de Humanidades Digitales*, *6*, 122-128. https://doi.org/10.5944/rhd.vol.6.2021.32297

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities, *Journal of Digital Humanities*, *2*(1), 49-66.

SEPLN (2020). *Informe SEPLN 2020: Hacia una estrategia para la IA centrada en las tecnologías del lenguaje en España*. http://www.sepln.org/actualidad/noticias/publicacion-de-la-estrategia-de-procesamiento-del-lenguaje-natural

Spence, P. (2014). La investigación humanística en la era digital: mundo académico y nuevos públicos. *Janus Digital*, Annex *2*, 117-131.

Tuzzi, A., & Cortelazzo, M. (2018). What is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer. *Digital Scholarship in the Humanities*, *33*(3), 685-702.

Ueda, H., Sanchez-Prieto, P., & Moreno Sandoval, A. (2020). Lematización y visualización cartográfica del corpus CODEA. *Estudios de lingüística de español*, *42*, 245-261.

Vázquez, A., & García-Serrano, A. (2015.) Anotación y representación temporal de tweets multilingües. *Procesamiento del Lenguaje Natural*, *54*, 53-60.

Vicente-Díez M.T., Moreno-Schneider, J., & Martínez P. (2010). Temporal Information Needs in ResPubliQA: an Attempt to Improve Accuracy. The UC3M Participation CLEF 2010, LABs and Workshops. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.3558&rep=rep1&type=pdf

Vivó Capdevila, E.P. (2021). Modelizando una literatura en el olvido: LDA aplicado a corpus españoles sobre Guinea Ecuatorial y. Filipinas. *HDH 2021-Scire Vías. Humanidades Digitales y Conocimiento*. 4-8 octubre. España.

Webis Group (2021). *PAN is a Series of Scientific Events and Shared Tasks on Digital Text Forensics and Stylometry*. https://pan.webis.de/

Wille, R. (1992). Concept Lattices and Conceptual Knowledge Systems. *Computers & mathematics with applications*, *23*(6), 493-515.

Yang, T. I., Torget, A. J., & Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. *Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96-104. ACL.