Colección
Ramón Acuña García

Inteligencia artificial e incunables poéticos: un modelo de transcripción automática

Enrique Rípoll Cervera

Inteligencia artificial e incunables poéticos: un modelo de transcripción automática

Inteligencia artificial e incunables poéticos: un modelo de transcripción automática

Enrique Ripoll Cervera

ULTREIA
Editorial del IVEMIR-UCV
(Institut Isabel de Villena d'Estudis Medievals i Renaixentistes)

Universidad Católica de Valencia San Vicente Mártir
Colección Ramón Arnau García, 5

Universidad Católica de Valencia San Vicente Mártir
IVEMIR Institut Isabel de Villena d'Estudis Medievals i Renaixentistes

Valencia

2024

Título original: *Inteligencia artificial e incunables poéticos: un modelo de transcripción automática*
1ª Edición: 2024

Esta publicación es resultado del proyecto *Poesía, ecdótica e imprenta* (PID2021-123699NB-I00), financiado por:

Editorial del IVEMIR-UCV (Institut Isabel de Villena d'Estudis Medievals i Renaixentistes) de la Universidad Católica de Valencia San Vicente Mártir.

Servicio de Publicaciones
Calle Quevedo, 2
46001 Valencia, España
Telf. +34 963 637 412
Mail: editorial.ultreia@ucv.es

Colección: Ramón Arnau García
Editora y coordinadora de la obra: Anna Peirats
Diseño de portada: Imagen Institucional de la Universidad Católica de Valencia
Maquetación: Glaux Publicaciones Académicas
Impresión: Artes Gráficas Soler

ISBN: 978-84-128605-7-3
Depósito Legal: V-4631-2024

Imagen de la portada: Íñigo de Mendoza (1492): *Coplas de la vita Christi*. Biblioteca Nacional de España, INC/2900, f. III^{r} – 92VC ©Biblioteca Digital Hispánica. https://bdh-rd.bne.es
Impreso en España

Colección Ramón Arnau García

5

ÍNDICE

1. Introducción

Desde sus inicios más tempranos, las humanidades digitales se han desarrollado gracias al impulso ejercido por investigadores avezados en la informática, que han sido verdaderos pioneros, si no visionarios de su potencial de aplicación a las disciplinas humanísticas. Entonces se denominaban todavía *nuevas tecnologías* y, si acaso, se les añadía el complemento nominal de *aplicadas a las humanidades*. Hace más de una década, cuando ya no eran tan nuevas, algunos de estos humanistas sensibles a lo digital, normalmente voces autorizadas de la filología, insistieron en la necesidad de renombrarlas como *humanidades digitales*, precisamente, para reivindicar la atención que se le ha prestado desde las sensibilidades y necesidades de esta rama del conocimiento o, por lo que nos atañe aquí, de la disciplina filológica, en concreto. Las inquietudes personales de cada uno de estos primeros humanistas digitales fueron el caldo de cultivo que permitió que se desarrollaran sus investigaciones, entonces, en efecto, pioneras, pero que supusieron unos cimientos sólidos sobre los que crecer, desarrollarse y expandirse, en este caso unidos a la propia evolución social de lo digital. Hoy, prácticamente, no se concibe una investigación en humanidades y menos aún un proyecto financiado que no incorpore los recursos informáticos en su obtención, gestión o difusión de datos.

Ante la necesaria expansión del mundo digital en las humanidades, si bien es cierto que algunos de los perfiles de estos investigadores han cambiado, al menos en buena parte, también lo es que, incluso, los más sensibles e interesados por lo digital se limitan a la utilización de softwares

preconcebidos, si no es que acaban recurriendo a ingenieros informáticos cuando necesitan generar algunos personalizados ante el surgimiento de necesidades científicas específicas o porque la gestión y explotación de algunos de ellos requieren un apoyo o asesoramiento técnico de mayor calado. Es recurrente que, en reuniones científicas de filólogos, incluidas aquellas que tienen como tema central del encuentro las propias humanidades digitales, como he tenido oportunidad de constatar, emerja la necesidad y, paradójicamente, se encuentren ante la dificultad de colaborar con personal cuya formación de base pertenece a esta disciplina bien por su disponibilidad, bien por cuestiones de financiación, bien por el propio entendimiento entre disciplinas y objetivos diferentes. Todo ello es lo que ha llevado a muchos grupos de investigación a cierta autosuficiencia digital o a colaboraciones puntuales con estos ingenieros, siempre con las limitaciones que ello conlleva y, al mismo tiempo, con la riqueza que ofrece este ejercicio de interdisciplinariedad.

Si las circunstancias y las sensibilidades personales de ciertos humanistas y, para el caso que nos ocupa, de esos filólogos pioneros en lo digital, fueron las que determinaron su acercamiento a la informática, sin ser especialistas estrictos en esta materia, con esta monografía vengo a explorar el camino inverso de tal interdisciplinariedad, en una dirección o sentido poco habitual en las humanidades digitales. Mi perfil me ofrece la oportunidad y el reto de explorar este campo con la solidez de unos conocimientos técnicos, que se han completado a lo largo de estos años y, en especial, durante el desarrollo de esta investigación, en cuanto a la historia de las humanidades digitales y de su aplicación filológica, delimitando sus principales campos de actuación, para desembocar en el principal objetivo de esta investigación: la generación de un modelo extendido para la transcripción automática de incunables poéticos. Esta era una necesidad de este campo de estudio, porque los desarrollos que, en este sentido, se habían hecho desde la filología se habían aplicado, sobre todo, a textos en prosa e impresos del siglo xvi y, más recientemente, a cierta exploración en la incunabilística, pero en ningún caso en un proyecto que considerase las peculiaridades de los impresos poéticos, cuyos rasgos de *mise en page* no solo los caracterizan, sino que dificultan su reconocimiento automático. A tal especificidad, si no dificultad o reto, añadimos otra, como es la antigüedad de los materiales, optando, de manera decidida, por los incunables poéticos, puesto que, entrenándose con ellos, la inteligencia artificial facilitaría el reconocimiento de impresos con tipografía gótica de épocas posteriores, con más ejemplares, a menudo en mejor estado y con menor inestabilidad tipográfica.

Las primeras versiones del software de OCR (*Optical Character Recognition*) únicamente permitían reconocer un pequeño conjunto de tipografías

hasta que, en 1974, Ray Kurzweil comercializó uno de carácter universal capaz de identificar la mayor parte de tipografías contemporáneas (Craine, 2022). Actualmente, el OCR se puede llevar a cabo con un porcentaje de error cercano a cero, aunque depende de las imágenes de las que se parta (Terras, 2012). De hecho, *Gallica*, la biblioteca digital de la Bibliothèque nationale de France, lo ha aplicado a algunos de sus facsímiles digitales con dispar resultado, debido a que el porcentaje de error se incrementa conforme aumenta la antigüedad de la obra original. A partir de unas pruebas en su buscador con diferentes ediciones de obras de los siglos XIX y XX, obtenemos un porcentaje estimado de acierto del 99,99%, mientras que, si escogemos una del siglo XV, cae hasta el 66,61%, lo que se traduce en que, aproximadamente, se transcribirán mal 34 palabras de cada 100 o, lo que es lo mismo, 272 palabras incorrectas por página, si consideramos que una página con letra estándar de tamaño 11, que es la que ofrecen, por defecto, los procesadores de texto, tiene 800 palabras. Weichselbaumer considera que la transcripción automática resulta de utilidad en el caso de impresos posteriores a 1800 (2020, p. 75), pero que, aplicándola a obras anteriores al siglo XIX, el porcentaje de error es tan elevado, que el resultado deja de tener utilidad; no obstante, de este mismo ejercicio con *Gallica*, obtenemos un 98,94% en el caso de impresos del siglo XVIII, por lo que, en realidad, deberíamos matizar sus conclusiones y avanzar la fecha hasta 1700. Ahora bien, no debemos desatender para estos datos que, en el caso de *Gallica*, se utiliza un OCR clásico y no especializado en materiales antiguos.

Frente a ello, el sucesivo aumento de la potencia de cálculo de los ordenadores ha provocado, en los últimos años, el resurgimiento de la inteligencia artificial (IA) y, en concreto, la aplicación de redes neuronales artificiales (ANN): la unión de una multitud de elementos de computación, a los que se le conoce como neuronas artificiales, que se organizan en una red con interconexiones y que simulan, en cierta manera, la forma de trabajo de las neuronas de los mamíferos. Aunque el nacimiento de las neuronas artificiales se suele establecer a principios de 1940, a raíz del trabajo de MacCulloch y Pitts (1943), que propusieron unos dispositivos con umbrales de funcionamiento binario, la primera aplicación práctica documentada en el campo del OCR es la de LeCun *et al.* (1989) para el reconocimiento automático de códigos postales de EEUU manuscritos. Para ello, utilizó redes neuronales a las que alimentó con imágenes clasificadas para que aprendiesen a reconocer los dígitos de forma autónoma.

En efecto, las ANN posibilitan que una máquina aprenda y sea capaz de construir un modelo en base a un conjunto de muestras representativas, que permitirá clasificar otras nuevas sin intervención humana. Dentro del ámbito de la filología, se han aplicado, muy recientemente, a la

transcripción automática de manuscritos (De Sousa Neto *et al.*, 2020) y de impresos antiguos, de lo que son ejemplo los resultados ofrecidos por el *Progetto Mambrino*, con sede en la Università degli Studi di Verona, que ha construido un modelo que permite el reconocimiento de textos del Siglo de Oro con una tasa de error cercana al 1% (Bazzaco, 2020, p. 551). Sobre tales resultados, basados en impresos en prosa de los siglos xvi y xvii, se establece la hipótesis sobre la que se construye esta monografía: la posibilidad de desarrollar con éxito la transcripción automática de incunables poéticos, cuyas peculiaridades generan un reto complejo e inexplorado hasta esta investigación.

La poesía impresa, a diferencia de la prosa, ofrece líneas de texto de diferente extensión, por las medidas de los versos, no solo cuando cohabitan el arte menor y mayor, sino por la falta de identificación sistemática entre un pie métrico y un número determinado de grafías y/o tipos, lo que provoca una ausencia de justificación completa en el margen derecho de la caja tipográfica de cada columna; a esto habría que sumar el alto número de espacios interestróficos, con una casuística exponencialmente mayor que la que supone el cambio de capítulos de una edición en prosa, además de que tales blancos no siempre se alinean entre columnas por la dificultad de la puesta en página de las coplas. Esta voluntad de incorporar completas todas las estrofas de una columna la encontraremos, de hecho, en los dos incunables más antiguos de las *Coplas de la vita Christi*, que Brian Dutton (1990-1991) identificó como 82IM y 82*IM, respectivamente. Por otro lado, la antigüedad de los testimonios añade nuevos problemas al reconocimiento de imágenes y transcripción automática mediante inteligencia artificial: a) la propia tipografía es mucho más inestable en el siglo xv que en los siglos xvi y xvii, tanto en su morfología, como en su aplicación; b) es más frecuente que las ediciones se conserven en ejemplares únicos, de manera que no se puede elegir entre digitalizaciones de mayor o menor calidad fotográfica, de lo que es paradigmática la reproducción borrosa que ofrece la Biblioteca Nacional de España del ejemplar único de 82IM, o descartar otros con mucho intervencionismo manuscrito o con errores de digitalización, como las numerosas duplicaciones de páginas en la reproducción de 83*IM, también de la BNE, que se puede suplir con el ejemplar de la British Library, por ejemplo; c) y, finalmente, tal estado de conservación desemboca en mutilaciones de los ejemplares, que llegan a su máximo exponente en el caso del cancionero incunable 99VC, del que se conserva también un ejemplar único en The Morgan Library & Museum, con importantes problemas de reconocimiento tipográfico, a pesar de la calidad de la reproducción fotográfica, por los graves daños materiales en prácticamente todas sus hojas.

Partiendo de la hipótesis de que es posible el reconocimiento automático de caracteres de impresos incunables poéticos con unas tasas de éxito del 100% o cercanas a este porcentaje, se establece el objetivo principal de esta investigación, que es, como explicita su título, generar un modelo extendido con este fin a partir de un OCR especializado. El punto de partida será la evaluación comparativa de los principales softwares disponibles en el mercado, en cada uno de los cuales se creará un modelo de transcripción aplicado únicamente a un ejemplar de una edición incunable, con el fin último de elegir el más adecuado a este cometido.

Para generar un modelo extendido de esta inteligencia artificial, se debe partir, necesariamente, de un muestrario tipográfico representativo del periodo elegido. Las *Coplas de la vita Christi* de fray Íñigo de Mendoza, identificada por Dutton con el ID 0269 (1990-1991, VII, 21), es la obra poética de la que más ediciones incunables conocemos, hasta ocho, que no solo abarcan, prácticamente, todo el periodo del siglo xv en que encontramos poesía de cancionero impresa, desde 1482 hasta 1499, sino que también ofrecen una muestra muy representativa de talleres de imprenta dispersos por la geografía española (Zamora, Zaragoza, Burgos y Sevilla), todos ellos, además, productores de otros incunables poéticos. Estos ocho impresos nos ofrecen un corpus de estudio con un total de diez tipografías diferentes, lo que, en definitiva, nos permitirá entrenar un modelo extendido capaz de transcribir otros incunables poéticos —e, incluso, impresos posteriores en gótica— salidos no solo de estos mismos talleres, sino de otros diferentes, como objetivo último de esta monografía.

2. De la informática computacional a las humanidades digitales

Actualmente, ya no concebimos la investigación sin la participación de la informática. Ya sea desde el uso más simple que le podamos dar al procesador de textos para ir volcando nuestros avances, hasta el empleo de complejos algoritmos que permiten tratar volúmenes de datos inabarcables de forma manual, el uso de aplicaciones y algoritmos se ha convertido en nuestro día a día y, por supuesto, la investigación en humanidades no es ajena a ello. Tal como apunta Lucía Megías, en el estado en el que nos encontramos, ya no se trata tanto de evaluar cómo influirá en las humanidades en un futuro, más o menos cercano, más o menos lejano, una tecnología —que ya no podemos seguir considerando *nueva*—, sino de aceptar cómo esta ya ha modificado aspectos básicos de nuestra vida cotidiana y profesional (Lucía Megías, 2003, p. 91).

El término *digital* ahora acompaña a parte de la terminología clásica: tenemos bibliotecas digitales, libros digitales y, por supuesto, humanidades digitales. Incluso se va más allá y algunos términos de uso habitual se presuponen de la familia *digital*, sin precisar con ello mayor especificación, como le ocurre al *documento*: si solicitamos que nos envíen uno, esperamos recibir un correo electrónico con un adjunto, no una carpeta de folios manuscritos. Sin embargo, para llegar a este punto en el que nos encontramos, ha tenido que pasar más de un siglo, un tiempo en el que nos hemos ido adaptando y evolucionando de manera conjunta en una simbiosis artificial entre la humanidad y las máquinas.

2.1. El primer ordenador y su programadora

El nacimiento de la informática no ha ido unido a las ciencias humanas, sino a las experimentales. La aplicación de las máquinas a la automatización de los cálculos se sitúa en el siglo xvii cuando Blaise Pascal —matemático, físico, teólogo y filósofo francés— hizo su gran contribución a la ciencia con el diseño y la construcción de la primera calculadora mecánica, conocida coloquialmente como *Pascalina*[1]. Su padre era recaudador de impuestos y, ante la ausencia de una unificación monetaria en la Francia de aquella época[2], se veía obligado a realizar repetidos cálculos de conversión de monedas. Viendo Blaise la penosa tarea de su padre, "se le ocurrió imaginar un aparato que hiciera operaciones mecánicamente y poder liberar así a su progenitor de tener que realizar las operaciones a mano" (Gutiérrez Vázquez, 2012, pp. 108-109). Su primer modelo le llevó tres años de construcción, entre 1642 y 1645 y hacía básicamente las operaciones de suma y resta[3]. Dado el éxito que tuvo, estableció una estrategia comercial de venta en la que construía de forma personalizada cada una de las máquinas según los requerimientos del comprador, adaptando las operaciones que era capaz de hacer (Adamson, 1995, p. 233).

Wilhelm Gottfried Leibniz, filósofo y matemático alemán, conoció las calculadoras de Pascal en una estancia en París a principios de 1670. Estudió su funcionamiento y observó que tenían determinadas limitaciones que, con otro diseño, se podrían solucionar. Se puso manos a la obra y entre 1672 y 1694 desarrolló una nueva calculadora mecánica que sumaba, restaba, multiplicaba, dividía y extraía raíces, la *step reckoner* (O'Regan, 2016, pp. 38-39). Fue la primera calculadora que podía realizar las cuatro operaciones aritméticas fundamentales. Con un funcionamiento basado en un cilindro dentado con una rueda acoplada que hacía el conteo, su ingenioso mecanismo ha sido la base de funcionamiento de las calculadoras hasta el comienzo de la era digital en la segunda mitad del siglo xx. Sin embargo, su gran aportación al mundo de la informática se produjo

---

1 Aunque se estima que Pascal construyó unas cincuenta máquinas, se tiene constancia de la existencia de nueve, siete de ellas en museos públicos de Francia y Alemania, y dos en colecciones privadas (Rojas-Sola *et al.*, 2021, pp. 1-2).

2 "In the seventeenth century, the currency in circulation in France was the livre, which was equivalent to the franc. The division and subdivision of this currency were the sous (sols) and the denarii (deniers), whereby one livre was equivalent to 20 sous, and the sou to 12 denarii" (Rojas-Sola *et al.*, 2021, p. 1).

3 Aunque también era capaz de multiplicar y dividir, estas operaciones las hacia apoyándose en la suma y la resta, aplicadas de forma consecutiva.

en 1703, año en el que publicó un artículo sobre el sistema de numeración binario. En él, detalló de forma rigurosa su funcionamiento con ejemplos de sumas, restas, multiplicaciones y divisiones (Leibniz, 1703, p. 86). Este sistema, dada su simplicidad al emplear únicamente dos dígitos en lugar de los diez del sistema decimal que utilizamos normalmente, se tomó como base para el funcionamiento interno de los ordenadores digitales.

El siglo XIX, con la revolución industrial, propició otra gran mejora de las máquinas calculadoras. En concreto, el 14 de junio de 1822, el matemático británico Charles Babbage sembró la semilla de la informática cuando presentó ante la Astronomical Society of London su *Máquina Diferencial*[4], un artilugio "which by the application of a moving force may calculate any tables that may be required" (Babbage, 1825, p. 309). Una proeza de la ingeniería que, dada una función matemática, mediante el movimiento sucesivo de una palanca, imprimía la secuencia numérica de la salida de una función en un tambor de cera. No obstante, ni era una calculadora, ya que no realizaba operaciones aritméticas simples entre números, ni era un ordenador, al no ser programable. En resumen, su diseño no permitía hacer otra cosa distinta al cálculo de funciones. Sin embargo, la inquietud de Babbage hizo que continuara mejorando sus inventos y, en 1837, propuso la *Máquina Analítica*. Aunque seguía siendo un dispositivo mecánico, es decir, funcionaba a base de manivelas que movían engranajes, sin utilizar electricidad, en esta ocasión, sí que era de propósito general. Por tanto, no se limitaba a hacer una única tarea: podía programarse a voluntad, lo que significaba que era capaz de efectuar cualquier cálculo si se establecía la configuración adecuada. Aunque los entresijos del funcionamiento de su nueva máquina eran teóricos y únicamente estaban esbozados en papel, eso no impidió a su creador mostrar "the degree of assistance which mathematical science is capable of receiving from mechanism" (Babbage, 1837, p. 19).

Babbage no dudaba de que sería un éxito y no cejaba en su empeño de promocionarla. Organizaba los fines de semana fiestas en las que reunía a diversos personajes del entorno cultural, científico y aristocrático de la época.

---

4 El London Science Museum construyó una réplica de la segunda versión del motor diferencial en 1991, con motivo del bicentenario del nacimiento de Babbage. Se puede ver en línea en https://collection.sciencemuseumgroup.org.uk/objects/co526657/difference-engine-no-2-designed-by-charles-babbage-built-by-science-museum-difference-engine [consulta 17/10/2022].

Fue en una de estas celebraciones donde Lady Byron y su hija Ada[5] se vieron cautivadas por sus ideas e inventos, como se atestigua en una de sus cartas:

> Fuimos las dos a ver la máquina pensante (que es lo que parece) el lunes pasado. Elevó diversos números a la segunda y tercera potencia, y extrajo la raíz de una ecuación cuadrada. Apenas conseguí una débil noción de los principios en los que se basa su funcionamiento[6].

La industrialización del siglo xix produjo cambios sociales y laborales, dado que "las máquinas transformaban, en efecto, el modo de producir los bienes y de hacer las cosas, pero también —lo que era más importante— estimulaban la imaginación de la gente sobre cómo podían hacerse las cosas" (Essinger, 2015, p. 58). En este ambiente de revolución, la joven Ada se sentía atraída por los ingenios que automatizaban los trabajos —uno de los que más le sorprendían era el *telar de Jacquard*[7], que permitía tejer intrincados diseños sin apenas intervención humana y que Babbage había tomado como inspiración—, así como por las leyes matemáticas que los regían. Es por ello que decidió buscar un tutor que la instruyera.

> En el verano de 1840, lady Byron le encontró uno: el famoso matemático y lógico Augustus De Morgan, que había estudiado en el Trinity College de Cambridge y tenía amistad con Babbage. Con su ayuda, Ada progresó rápidamente en el estudio de la disciplina

---

5 Su nombre completo era Ada King, condesa de Lovelace y actualmente conocida como Ada Lovelace. Fue la única hija del matrimonio entre el poeta Lord Byron y Anna Isabella Noel Byron, que se separaron dos meses después de su nacimiento. Lord Byron, al que no conoció personalmente, abandonó Gran Bretaña para siempre; sin embargo, plasmó su pesar por la separación de su hija en el canto 3, estrofa 1 del poema *Las peregrinaciones de Childe Harold*: "¿Es tu rostro como el de tu madre, mi preciosa niña, / Ada, única hija de mi casa y de mi corazón? / La última vez que vi tus jóvenes ojos azules, sonreían, / y entonces nos separamos, —no como nos separamos ahora, / sino con esperanza—" (Hollings, 2019, p. 15).

6 Correspondencia de Lady Byron al doctor King (Dep. Lovelace Byron, 77, f. 127ʳ), *apud* Hollings *et al.*, 2019, p. 61.

7 El *telar de Jaquard*, patentado en 1804 por Jean-Charles Jacquard, teje un diseño a partir de la posición de los agujeros de unas tarjetas que estimulan unos resortes conectados a las agujas (Keats, 2011, p. 9). Otro invento con un funcionamiento similar es la pianola, pero en este caso, reproduce una melodía.

> que más le interesaba. Por primera vez, al parecer, se sintió plenamente satisfecha en lo intelectual. (Essinger, 2015, p. 102)

Ada siguió con interés los trabajos de Babbage y en su traducción del artículo original de Menabrea sobre la *Máquina Analítica*[8], aprovechó su formación matemática para desarrollar ciertos apéndices, en uno de los cuales creaba un algoritmo que calculaba los números de Bernoulli adaptado al dispositivo (Menabrea, 1843, pp. 722-731): la máquina se había programado. Se demostraba que se podía adaptar a distintos cálculos variando la posición de palancas y engranajes. Babbage utilizó este artículo en su autobiografía para defender su universalidad y crear su famosa cita sobre las capacidades de las máquinas:

> These two memoirs taken together furnish, to those who are capable of understanding the reasoning, a complete demonstration — *That the whole of the developments and operations of analysis are now capable of being executed by machinery*. (Babbage, 1864, p. 136)

Con ella venía a decirnos, en última instancia, que, a partir de ese momento, las matemáticas y las máquinas caminarían juntas y de la mano.

Aunque Charles Babbage finalmente no la vio construida, la historia de su *Máquina Analítica* continuó a principios del siglo xx, de la mano de su hijo Henry P. Babbage, que utilizó los planos de su padre para desarrollar una calculadora capaz de sumar, restar, multiplicar y dividir[9]. Trabajaba internamente con números decimales, es decir, con los dígitos del 0 al 9, y no en binario, esto es, con solo los dígitos 0 y 1, a la manera de los ordenadores actuales. Esa característica aumentaba considerablemente la complejidad de su construcción, por lo que tan solo se desarrolló lo que se podría considerar el procesador central, *the mill*, no la máquina completa. Sin embargo, hoy en día siguen sin estar claras las razones por las que finalmente el propio Babbage no construyó el dispositivo.

---

8 Dado que Babbage no consiguió convencer al gobierno británico para que financiase la construcción de la máquina analítica, buscó ayuda externa y difundió su propuesta en el extranjero. A una de esas conferencias, en este caso en Italia, asistió Luigi Federico Menabrea, quien, seducido por la idea de Babbage, escribió en 1842 su artículo en francés.

9 Está en la colección del Science Museum Group y se puede ver una descripción en https://collection.sciencemuseumgroup.org.uk/objects/co62246/henry-babbages-analytical-engine-mill-1910-analytical-engine-mills [consulta: 19/04/2023].

> A number of Babbage's biographers have argued that he could not build his engines because British machine technology was not up to the task. Reconstruction of his early machines has led some to argue that the level of British technology employed was more advanced. Lindgren and Bromley's analyses, based upon their historical and actual reconstructions suggest that while the technology was capable of producing Babbage's engines, his engineering skills were not. (Seidel, 2000, p. 42)

Independientemente de que la razón de ello fuesen los medios de los que se disponía o sus capacidades constructivas, tal como Babbage cita a lord Byron en sus memorias para dar inicio al capítulo de la *Máquina Analítica* con "Man wrongs, and Time avenges" (Babbage, 1864, p. 112)[10], el tiempo lo pone todo en su sitio y, actualmente, este dispositivo se considera el primer ordenador, el algoritmo para calcular los números de Bernoulli se interpreta como el primer programa de computador y Ada Lovelace se reconoce como la primera programadora informática.

10 Babbage extrae la cita de *The Prophecy of Dante* de lord Byron.

2.2. El ordenador electrónico

Durante los primeros años del siglo xx algunos inventores se inspiraron en el trabajo de Babbage para diseñar sus máquinas analíticas, como Ludgate en Irlanda en 1909, Torres Quevedo en España en 1920, Couffignal en Francia en los años 30 y Bush en la misma época, en Estados Unidos, con su *analizador diferencial analógico* (Ralston & Reilly, 1983, p. 533); sin embargo, la verdadera revolución informática se generó en el marco de la II Guerra Mundial.

En 1936, en medio de una Alemania prebélica, un ingeniero llamado Konrad Zuse, cansado de los tediosos cálculos numéricos que hacía diariamente en su puesto de trabajo, creó el *Z1* en el salón de la casa de sus padres, "un aparato casi totalmente mecánico, capaz de ejecutar las cuatro operaciones aritméticas (suma, resta, multiplicación y división) en cualquier secuencia y con números almacenados en una memoria" (Rojas, 1997a, p. 22), pero que, además, era programable, ya que leía la secuencia de los cálculos que debía efectuar a través de una cinta perforada que se introducía desde el exterior.

El *Z1* fue construido de forma artesanal con finas láminas de metal cortadas mediante una sierra de calar por Zuse y sus amigos. Estas piezas actuaban a modo de interruptor que indicaba dos posiciones: *abierto* o *0* y *cerrado* o *1*, es decir, funcionaba con números binarios. Sin embargo, pese a la ardua tarea manual que supuso crear y ensamblar todo el conjunto, la máquina no operaba de forma fiable[11]. En vista de los problemas que se derivaban de su funcionamiento, su creador replanteó el diseño y optó por sustituir las complejas piezas mecánicas por relés electromagnéticos[12]. Zuse finalizó su nueva creación que enmendaba los errores del *Z1* en 1941 y lo llamó *Z3*[13], un dispositivo considerado "the first machine in the world that could be said to be a fully working computer with automatic control of its operation"

---

11 El *Z1* fue destruido por un bombardeo durante la II Guerra Mundial, pero en 1989, el Deutsches Museum finalizó una réplica bajo la supervisión del propio Konrad Zuse. En esta ocasión se utilizaron instrumentos precisos para efectuar los cortes y, pese a ello, también se encalla como la máquina original (Rojas, 1997a, p. 23).

12 Los *relés electromagnéticos* funcionan a la manera de los interruptores, pero accionados mediante electricidad en lugar de manual o mecánicamente.

13 El *Z3*, al igual que el *Z1*, también fue destruido en la II Guerra Mundial. Zuse lo construyó nuevamente en 1967 y lo donó al Deutsches Museum, aunque no documentó su arquitectura interna ni su funcionamiento, por lo que únicamente era capaz de operarlo él. Entre 1999 y 2001, un grupo de investigadores generó una

(Williams, 1985, p. 222). Aunque seguía sin ser un ordenador electrónico, era electromecánico.

En el mismo año en el que Zuse creaba el *Z1* en Alemania, el matemático inglés Alan Turing escribía las bases de las ciencias de la computación en Gran Bretaña con la descripción de su máquina universal, también llamada *Máquina de Turing*, un modelo matemático que sirve para demostrar que existen problemas irresolubles por las máquinas y que define el funcionamiento de los ordenadores (Turing, 1937, pp. 232-233). El novedoso enfoque de su artículo le valió para que le invitaran a una estancia en Princeton, considerado en aquel momento el lugar de encuentro de los grandes matemáticos. Allí se interesó por la criptografía y "construyó un multiplicador electrónico capaz de cifrar mensajes por medio de la multiplicación de grandes números binarios entre sí" (Leavitt, 2006, p. 152)[14]. Su conocimiento de las máquinas y la criptografía le valieron para que el gobierno británico le solicitase, a su vuelta de Princeton, su incorporación al equipo de criptógrafos que trabajaban en la desencriptación de los mensajes emitidos por los alemanes.

Las comunicaciones a larga distancia se efectuaban mediante ondas de radio que viajaban por el aire, por lo que cualquiera podía interceptarlas y escucharlas. Esto, en tiempos de guerra, era inasumible. El enemigo no podía saber de antemano las órdenes de movimiento de las tropas. Para evitarlo, se idearon distintas formas de encriptación de los mensajes. Es por ello que, durante la Guerra Civil Española, "the Nationalists used the *Enigma* machines as did the Germans from the Condor Legion and the Italians for their naval communications" (Soler Fuensanta, 2004, p. 266)[15]. De apariencia similar a una máquina de escribir, la particularidad de *Enigma* y su gran valor residía en que, con la ayuda de unos engranajes, cables y rotores que cambiaban de posición cada vez que se pulsaba una tecla, conseguía que el mensaje que se introducía se transformase en otro

---

nueva réplica (Rojas *et al.*, 2005, pp. 28-30) con la ayuda de la documentación que había hecho Rojas sobre el funcionamiento interno del *Z1* y del *Z3* (1997b, pp. 14-15), aunque esta vez se simularon diversas partes del sistema mediante software. En 2008, Horst Zuse construyó una nueva versión del *Z3* sin componentes virtuales, pero empleando componentes electrónicos modernos (2013, pp. 285-287).

14 El principio de la multiplicación de dos grandes números entre sí es la base de la criptografía moderna. Se utiliza para encriptar los datos y las comunicaciones por Internet, aunque se eligen dos números primos. La firma digital también se crea de esta forma.

15 Para esta cuestión, véase Rejewski, 1981, pp. 216-224.

totalmente distinto y únicamente interpretable por otra máquina *Enigma* que tuviese los engranajes y el cableado en la misma posición inicial. De forma simplificada: si se introducía la letra *T*, podía sacar la letra *G*, pero si, a continuación, se introducía de nuevo la letra *T*, generaba otra letra distinta. El número de combinaciones que conseguía hacía imposible que los mensajes pudiesen desencriptarse de forma manual en un tiempo razonable (Rejewski, 1981, pp. 214-216). No obstante, aquí residía su talón de Aquiles. Aunque los cálculos efectuados por un ser humano eran lentos, se sabía que una máquina podía acelerarlos; por tanto, la gran incógnita era si se podía construir una máquina adaptada a tal fin.

Turing tenía el perfil perfecto para ese trabajo. Era matemático, criptógrafo y diseñador de máquinas de computar. Él, junto a un equipo multidisciplinar en el que no faltaban los matemáticos y los lingüistas, se fueron a Bletchley Park, una instalación militar secreta situada entre Cambridge y Oxford, camuflada bajo la apariencia de una mansión victoriana rodeada de pabellones que simulaban cobertizos de caza. Su emplazamiento, alejada de las grandes urbes, la hacía perfecta para que el movimiento de gente no levantase sospechas. Allí, en la primavera de 1940, Turing creó la primera máquina llamada *Bomba*[16] a la que le siguieron varias más que convirtieron a Betchley Park en un gran centro de desencriptación de mensajes militares (Copeland, 2012, pp. 63-64). Las *Bombas* eran unos dispositivos electromecánicos que mejoraban los creados por el matemático polaco Marian Rejewski, los cuales se habían quedado obsoletos ante una actualización de las máquinas *Enigma*[17]. Equipadas con multitud de rotores, cada una era capaz de simular el funcionamiento de varias máquinas *Enigma* simultáneamente, consiguiendo con ello desencriptar los mensajes en menos de un día, el periodo de tiempo máximo para que estos fuesen aprovechables.

Las *Bombas*, por descontado, fueron un gran éxito, pero su diseño no permitía programarlas, puesto que eran unos dispositivos construidos específicamente para un único propósito: buscar, en el mínimo tiempo posible, la combinación de rotores y cableado que tenía la máquina *Enigma* que había transmitido cierto mensaje. Su especialización comportó que,

16 No se tiene claro el motivo por el cual se le atribuyó el nombre de *Bomba*. Se baraja que fue por el sonido de tic-tac que hacía al funcionar o por un tipo de helado polaco que Rejewski, su primer inventor, comía (Leavitt, 2006, p. 167).

17 Entre 1938 y 1939 las máquinas *Enigma* sufrieron varias actualizaciones que complicaron el cálculo de las combinaciones posibles; una de las más importantes fue el incremento del número de rotores (Rejewski, 1981, p. 227).

en 1940, cuando los alemanes incorporaron en algunas transmisiones unas nuevas máquinas creadas por una empresa llamada Lorenz para encriptar mensajes, el bando aliado se quedase, de nuevo, sin posibilidad de saber su contenido.

Sin embargo, en Betchley Park, lejos de desistir, continuaron capturando y analizando las transmisiones de las máquinas *Lorenz* hasta que, finalmente, el 30 de agosto de 1941, un error humano[18] durante una de esas transmisiones abrió una brecha en su sistema que posibilitó la rotura del cifrado. A partir de ese momento, todo se aceleró. Bill Tutte, un joven químico recién graduado, consiguió replicar la máquina *Lorenz* y el matemático Max Newman, mentor de Turing, crear un prototipo de máquina electrónica que encontraba la configuración que se había utilizado para la transmisión de un mensaje. Max Newman le pasó el testigo a Tommy Flowers, un brillante ingeniero electrónico, para que mejorase su prototipo y Flowers no solo lo hizo, sino que creó el primer ordenador electrónico programable: el *Mark 1 Colossus*. En enero de 1944 se ponía en funcionamiento en Betchley Park esa nueva máquina que era capaz de romper el código Lorenz en horas y que se podía programar cambiando de posición una serie de interruptores y cables. Al *Mark 1* le siguió al año siguiente el *Mark 2*. En total, se estima que al final de la guerra había diez de ellos funcionando (Sale, 2000, p. 218). La información que se obtuvo gracias a ellos fue de vital importancia para el éxito de los aliados. El desembarco de Normandía no hubiese sido posible sin ellos.

> It provided vital information for the Normandy landings, and it confirmed that Hitler had been successfully misled by Allied disinformation into believing that the Normandy landings were to be a diversionary tactic. Further, it confirmed that no additional German troops were to be moved there. (O'Regan, 2021, p. 62)

18 Se produjo una primera transmisión en la que el receptor indicó que no había recibido correctamente el mensaje. El emisor lo volvió a enviar, pero sin cambiar la configuración de la máquina, lo cual era un error. Además, abrevió algunas palabras del mensaje original con el fin de ahorrar tiempo, lo cual era aún más grave al no haber modificado la configuración de la máquina. Estos dos descuidos unidos posibilitaron el análisis del sistema de encriptación que utilizaban las máquinas *Lorenz* para, finalmente, romperlo (Sale, 2000, p. 217).

Desgraciadamente, estos proyectos se mantuvieron en secreto y todos ellos fueron desmantelados[19] entre el final de la guerra y 1960, un hecho que, posiblemente, enlenteció la investigación y el desarrollo informático británico.

Estados Unidos, no obstante, tampoco fue ajeno a la innovación tecnológica que suponían los primeros ordenadores y, en el mismo año en el que Zuse comenzaba a construir su *Z1*, Howard Aiken, ingeniero eléctrico, influenciado por las ideas de Babbage, también comenzaba el desarrollo de su serie de máquinas *Mark* con la ayuda de IBM y la marina estadounidense (Williams, 1985, p. 241). El primer dispositivo, el *Harvard Mark I*, considerado también como uno de los primeros ordenadores electromecánicos,

> gave the world of scientists and engineers a visible proof that a complex machine could solve complicated mathematical problems by being programmed to execute a series of controlled operations in a predetermined sequence –and do so without error (Cohen, 2000, p. 107).

Su construcción comenzó en 1936, finalizó siete años después y permaneció en funcionamiento hasta 1959. Durante sus años de servicio se dedicó principalmente a cálculos militares hasta el final de la guerra (O'Regan, 2016, p. 58). En plena efervescencia de la II Guerra Mundial, los ordenadores suponían una ayuda inestimable. En este caso concreto, su uso en el campo balístico para llevar a cabo el cálculo de la trayectoria de misiles, sin la necesidad de hacer complejos cálculos de forma manual y sin errores, marcaba una clara ventaja. Dadas las circunstancias, la financiación no era un problema si con su construcción se obtenía una ganancia en el campo de batalla.

Aunque el *Mark I* fue todo un éxito, seguía siendo electromecánico y su construcción se dilató durante varios años, suficientes para que otros investigadores continuaran innovando. Era el momento propicio, ya que se unían los conocimientos en computación que se poseían, la experiencia en la construcción de máquinas automáticas y las grandes posibilidades de obtener financiación por la necesidad acuciante de acelerar los cálculos que tenían los países implicados en el conflicto bélico. Y fue justo cuando

---

19 En The National Museum of Computing de Betchley Park hay una habitación dedicada exclusivamente a alojar una reconstrucción funcional de un *Colossus*.

comenzaban a sonar los tambores de guerra en Europa, poco después de iniciarse la construcción del ordenador de Aiken, el también ingeniero eléctrico John Atanasoff y el recién graduado Clifford Berry establecían las bases de los ordenadores electrónicos con el *ABC*[20] en el Iowa State College[21]. Después de verificar sus ideas sobre unas pequeñas placas de prototipos electrónicos en 1939, dieron inicio al desarrollo de una versión funcional a escala completa.

Desafortunadamente, en 1942, cuando la máquina estaba casi acabada, Atanasoff, que permanecía ajeno al conflicto, fue llamado a filas, se interrumpió su trabajo y el *ABC* nunca llegó a completarse (Williams, 1985, pp. 267-270). Además, al no estar dentro de la esfera bélica, cayó en el olvido. No obstante, el círculo de investigadores en computación conocía las innovaciones que aportaba, dado que Atanasoff y su ayudante Berry habían compartido sus experimentos con otros colegas, y en concreto con John W. Mauchly.

> In December of 1940, Mauchly met Atanasoff at a meeting of the American Association for the Advancement of Science. This led to Mauchly's visiting Iowa State College during the summer of 1941 and talking with Atanasoff and Berry about the calculator they then had under construction. This visit led to continued discussions with Eckert, and anyone else who would listen, about the use of electronics for performing calculations. (Williams, 1985, p. 274)

Las ideas de Atanasoff y Berry, tal como se demostró en un juicio que se celebró a principios de los 70, fueron aprovechadas por los ingenieros J. Presper Eckert y John W. Mauchly para construir el *ENIAC* entre 1943 y 1945 en la Moore School of Electrical Engineering de la University of Pennsylvania, considerado durante mucho tiempo el primer ordenador electrónico (Spiegel *et al.*, 2000, p. 121)[22].

Un año después de su conversación con Atanasoff, Mauchly, físico y profesor de la Moore School of Electrical Engineering, escribió que

---

20 La sigla por la que se le conoce responde a las iniciales de sus inventores*: Atanasoff Berry Computer*.

21 Actualmente es la Iowa State University.

22 El nombre de *ENIAC* proviene de las iniciales de *Electronic Numerical Integrator and Computer*.

> a great gain in the speed of the calculation can be obtained if the devices which are used employ electronic means for the performance of the calculation, because the speed of such devices can be made very much higher than that of any mechanical device. (Mauchly, 1982, p. 355)

Por tanto, consideraba que se podía obtener más velocidad de cálculo si se utilizaban componentes electrónicos, como los empleados por el *ABC*, en lugar de mecánicos, que eran aquellos a los que recurría el *Mark I*. Obviamente, todo lo que implicase acelerar los cálculos era bienvenido por el mando militar. De hecho, las universidades estaban enfrascadas en una búsqueda de soluciones y, aunque el *ABC*, que era casi funcional, pasó desapercibido, no ocurrió lo mismo con las ideas de Mauchly, quien, junto a su compañero Eckert, que dominaba la parte electrónica, obtuvieron la financiación necesaria para desarrollar una máquina electrónica inspirada en el analizador diferencial de Bush (Ralston & Reilly, 1983, p. 535). Finalizaron la construcción del primer ordenador electrónico plenamente funcional en 1946: el *ENIAC*. Esta máquina pesaba sobre las 30 toneladas y consumía 150 kW, el equivalente a unas 15 000 bombillas LED y que se dedicaría plenamente a fines militares (O'Regan, 2021, p. 57). Fue destronada como primer ordenador electrónico después de un sonado juicio en el que se dictaminó que había tomado ideas del *ABC*, al que se le pasó el testigo[23].

El *ENIAC* era un ordenador electrónico que funcionaba con tubos de vacío, lo equivalente a los transistores actuales y, además, de propósito general. Por tanto, permitía programarle las operaciones que tenía que realizar, aunque para ello había que cambiar interruptores y cableado. Esto conllevaba un trabajo previo de planificación que podía llevar semanas, pero, pese a lo tedioso del proceso, el paso estaba dado y el primer ordenador electrónico estaba operativo y demostrando su supremacía sobre las máquinas mecánicas.

El reinado del *ENIAC* duró poco, ya que tenía dos problemas: utilizaba la representación numérica decimal[24] y no la binaria, lo que complicaba

---

23 En el juicio de Honeywell Inc. contra Sperry Rand Corp. *et al.* entre 1971 y 1973, por el que se invalidó la patente de 1964 del *ENIAC* sobre el diseño de ordenadores digitales, se dictaminó que la patente no era válida dado que se había solicitado un año después de poner en marcha para uso público el *ENIAC* y que, además, su diseño se había derivado del *ABC*.

24 En la representación decimal, se expresan los números con 10 símbolos, los números de 0 al 9. En la binaria, se utilizan únicamente 2, el 0 y el 1.

su arquitectura y, además, para reprogramarlo, había que hacer complejos cambios estructurales. Sus creadores no tardaron en percatarse de estos inconvenientes y, mientras finalizaba su construcción, diseñaron, junto al matemático John von Neumann, su sucesor: el *EDVAC*[25]. Este nuevo ordenador, aunque tomaba como inspiración el *ENIAC*, era una máquina totalmente distinta. Ya no trabajaba en decimal, sino en binario, lo que simplificaba su diseño y, además, se le había incorporado una memoria en la que se almacenaban los programas y se evitaba la reconfiguración del cableado cuando se tenía que reprogramar[26]. Eckert y Mauchly tuvieron visión comercial y, viendo lo que tenían entre manos, abandonaron el proyecto antes de su finalización para crear su propia empresa, Eckert-Mauchly Corporation[27], con la que comercializaron el primer ordenador electrónico empresarial en 1951: el *UNIVAC I* (Ralston & Reilly, 1983, pp. 535-537)[28].

Con el *EDVAC* y su arquitectura, conocida como *Von Neumann* en homenaje a su creador y que aún permanece vigente en los computadores actuales, la invención del ordenador electrónico ya era una realidad. Y dado que la guerra también había terminado, era el momento de explorar las posibilidades de las computadoras más allá de los cálculos militares.

---

25 La sigla responde a *Electronic Discrete Variable Automatic Computer*.

26 John von Neumann escribió un documento titulado *First draft of a report on the EDVAC* fechado el 30 de junio de 1945, que se considera la primera descripción del funcionamiento de un ordenador con un programa almacenado en memoria (1945).

27 Poco después fue absorbida por Remington Rand, empresa dedicada a la fabricación de máquinas de escribir.

28 El primer ordenador de IBM fue el *701* y se creó en 1953, dos años después del *UNIVAC* (Fisher *et al.*, 1983, p. 17).

2.3. Primeras aplicaciones textuales

Los primeros ordenadores utilizaban un sistema común para la interacción con el exterior: las tarjetas perforadas. Eran unas cartulinas que contenían agujeros cuya disposición indicaba a la máquina los datos y las operaciones que tenía que hacer sobre ellos. Las tarjetas perforadas no eran una invención coetánea, ya las utilizaba el *telar de Jacquard* para puntear patrones a principios del siglo XIX. Asimismo, demostraron su potencial "processing information from the 1890 United States census" (Heide, 2009, p. 15), el primer gran proyecto en el que se almacenó información textual en ellas: los datos de los ciudadanos. Por tanto, por un lado, las tarjetas perforadas podían registrar números y texto, mientras que, por otro lado, los ordenadores eran capaces de leerlas y efectuar operaciones con ellas, una combinación que aprovechó el padre Busa para llevar a cabo su gran proyecto de humanidades digitales.

A finales de los años 40 y principios de los 50 del siglo XX, las computadoras eran una novedad y sus fabricantes se encontraban en una búsqueda constante de nuevas vías de uso de sus máquinas con el fin de aumentar las ventas. En ese momento de incertidumbre sobre las posibilidades que podían brindar, un sacerdote italiano, el padre Busa, desarrollaba su tesis sobre la obra en latín medieval de Santo Tomás de Aquino, escribiendo a mano en pequeñas cartulinas las concordancias de sus escritos[29]. La defendió en 1946 y la publicó en 1949, pero se propuso ir más allá y extender su estudio a todas las palabras de la obra completa y, para ello, tenía claro que iba a necesitar algún tipo de ayuda mecánica (Busa, 1980, p. 83), que vino por parte de IBM y de las tarjetas perforadas, que jugarían un papel fundamental en su desarrollo.

Aunque no son fáciles de explicar las razones[30] que llevaron a un fabricante como IBM, enfocado a las aplicaciones empresariales[31], a financiar un proyecto de humanidades digitales, ambos mantuvieron la colaboración durante 30 años hasta completar la lematización de las casi 10 600 000 palabras que forman el *Index Thomisticus* (Busa, 1980, p. 87).

> However, this proof of concept exercise used no computing and no programming. The main innovation was Busa's insight that

29 Escribió manualmente un total de 10 000 tarjetas que tenían un tamaño de 3"x 5" con todas las frases que contenían la preposición *in* o una palabra conectada con ella (Busa, 1980, p. 83).

30 "The Index Thomisticus was at once both Busa's research project and IBM public relations project" (Jacob, 2020, p. 9).

31 Su nombre, de hecho, proviene de las siglas de *International Business Machines*.

> commercial accounting machines could be used for humanities purposes with good results (Vanhoutte, 2013, p. 127).

No obstante, pese a esta crítica de que no se habían empleado ordenadores en los inicios del proyecto, pronto formaron parte de su ciclo de trabajo y jugaron un papel fundamental en la lectura y clasificación de las tarjetas (Busa, 1980, p. 85). Para introducirse en el ordenador, estas se debían puntear manualmente mediante unos dispositivos similares a las máquinas de escribir, que generaban el patrón de agujeros en la cartulina según la tecla que se pulsase. Fue de tal magnitud el trabajo de creación de esas tarjetas, que el padre Busa mantuvo una escuela de formación para operadores de máquinas perforadoras[32] desde 1956 hasta 1967 (Nyhan & Terras, 2017, p. 1). De esta manera, se podía incorporar a más gente al proyecto para absorber el ingente volumen de trabajo que suponía transferir las frases y las palabras de los escritos de Santo Tomás de Aquino a los códigos de las tarjetas perforadas. Finalmente, en 1980, se dio por concluido el proceso, con un total de 56 volúmenes impresos que habían comenzado a editarse seis años antes (Busa, 1974-1980). Este fue, por tanto, el primer gran proyecto de humanidades digitales, que, además, se fue adaptando progresivamente a la tecnología disponible en cada momento: comenzó con las tarjetas perforadas, pasó a utilizar cintas magnéticas (Busa, 1980, p. 85), se editó en papel, en CD-ROM (Donadío Maggi de Gandolfi, 1992, p. 233) y, actualmente, se encuentra disponible en línea[33]. Además, no ha caído en el olvido y su influencia llega hasta nuestros días, con proyectos de procesamiento de lenguaje natural que se basan en este corpus lingüístico, como el que se está desarrollando para conseguir que los ordenadores sean capaces de interpretar textos escritos en latín[34].

Aunque el padre Busa fue, posiblemente, el primero en crear concordancias con ayuda de una máquina, no fue el único. Cuando comenzaron a instalarse ordenadores en las sedes de las instituciones públicas y las grandes corporaciones, los dos principales fabricantes eran IBM y Remington Rand, que vendía el *UNIVAC I*, el primer ordenador electrónico comercial de los creadores del *ENIAC*. Ambas empresas mantenían una férrea

---

32 El alumnado de la escuela estaba compuesto en su mayor parte por mujeres que, con esa formación, podían obtener mejores oportunidades de empleo (Nyhan & Terras, 2017, pp. 1-2).

33 https://www.corpusthomisticum.org/it/index.age [consulta: 18/11/2022]. El contenido vertido en línea se ha extraído directamente de las cintas magnéticas originales (Alarcón, 2002, pp. 792-793).

34 El proyecto *Index Thomisticus Treebank* está etiquetando las palabras para que se puedan llevar a cabo análisis léxicos y sintácticos automáticos (Passarotti, 2019).

confrontación para intentar superarse una a la otra y, obviamente, si IBM estaba creando un catálogo textual, en Remington Rand no iban a quedarse atrás. En 1956, el reverendo John W. Ellison, bajo su auspicio, editaba el *Nelson's Complete Concordance to the Revised Standard Version Bible*[35].

> This work took nine months (800 000 words). The accuracy of the tapes was checked by punching the text a second time, on punched cards, then transferring this material to magnetic tape using a card-to-tape converter. The two sets of tapes were then compared for divergences by the computer and discrepancies *eliminated*. The computer output medium was also magnetic tape and this operated a Uniprinter which produced the manuscript sheets ready por typesetting. (Wisbey, 1965, p. 225)

Esto supuso toda una novedad en la forma de trabajar, dado que, en aquel momento, lo habitual era efectuar la entrada mediante tarjetas perforadas una única vez. Sin embargo, Ellison, para asegurarse de que los datos introducidos mediante el novedoso teclado y almacenados en las no menos innovadoras cintas magnéticas eran correctos, efectuó una segunda entrada mediante las ya probadas tarjetas perforadas, comparando después ambas versiones con ayuda del ordenador y consiguiendo, con este proceso, una doble validación.

No obstante, pese a los evidentes beneficios que se habían obtenido con la automatización, Ellison, en el prefacio de la obra, da cuenta de los dos problemas que se encontró al trabajar con una máquina. Por un lado, los primeros ordenadores se habían diseñado para utilizar el alfabeto latino y, por tanto, no eran capaces de trabajar con otros alfabetos, como el griego y el hebreo:

> The use of a computer imposed certain limitations upon the Concordance. Although it could be *exhaustive*, it could not be *analytical*; the context and location of each and every word could be listed, but not the Hebrew and Greek words from which they were translated. (Ellison, 1956)

---

35 En concreto, la versión americana *Revised Standard Version* (RSV) publicada en 1952 por la National Council of the Churches of Christ in the USA.

Por otro lado, no eran capaces de almacenar el conocimiento filológico de un ser humano y aplicarlo para tomar decisiones:

> A computer, at least in the present stage of engineering, can perform only the operations specified for it, but it will precisely and almost unerringly perform them. In previous concordances, each context was made up on the basis of a human judgment which took in untold familiarity with the text and almost unconscious decisions in grouping words into familiar phrases. This kind of human judgement could not be performed by the computer; it required a set of definite invariable rules for its operation. (Ellison, 1956)

Aunque el primero de los problemas está totalmente resuelto y, actualmente, los ordenadores son capaces de representar múltiples alfabetos, el segundo todavía está en proceso de superación. La inteligencia artificial y, en concreto, las redes neuronales han evolucionado en los últimos años y son capaces de aprender de forma automática; sin embargo, en determinadas situaciones, aún es necesaria la actuación de un experto que transmita su conocimiento para que el ordenador cree las reglas oportunas que lo representen.

La creación de concordancias fue todo un hito, pero las posibilidades del ordenador también se aprovecharon en otros campos filológicos; en concreto, en la traducción. El mismo año en que el padre Busa defendía su tesis, en 1946, Warren Weaver[36], biólogo, matemático e informático teórico, que había trabajado con ordenadores militares durante la II Guerra Mundial en Estados Unidos, mantenía una discusión con el cristalógrafo británico Andrew Donald Booth, que estaba de visita recorriendo las instalaciones informáticas. Durante esta conversación,

> Weaver suggested that all language might contain basic elements which could be detected by means of the techniques developed during World War II for the breaking of enemy codes, whereas Booth took the far more limited position that any digital computing

36 Colaboró con Claude E. Shannon en *The Mathematical Theory of Communication* (Shannon & Weaver, 1949), donde el primero estableció las bases matemáticas de la transmisión de datos y, por ello, se le conoce como el padre de una de las ramas de la informática: la teoría de la información. En la segunda parte del libro, Weaver ofreció una descripción introductoria de la teoría matemática que había desarrollado Shannon, avalándola.

> machine having the necessary storage capacity could make a dictionary translation. (Booth & Locke, 1955, p. 2)

Ambos eran conscientes de que los ordenadores se podían aprovechar para otros menesteres más allá de los cálculos matemáticos y, en concreto, que la traducción automática, al menos teóricamente, era uno de ellos. Booth regresó a su país y comenzó la investigación sobre el tema, pero su repercusión en Estados Unidos fue mínima hasta que, en 1949, Weaver escribió un memorándum que hizo de detonante. En él, plasmaba la correspondencia que había mantenido con diversas personas relacionadas con las ciencias de la computación y sus reflexiones sobre los problemas asociados con el hecho de utilizar una máquina para efectuar la traducción de un idioma a otro (Weaver, 1955, p. 16): "These problems included ambiguity of words, the semantic function of syntax, and the resolution of word order problems in different languages" (Vanhoutte, 2013, p. 124). El escrito de Weaver circuló rápidamente entre los investigadores de las distintas universidades y provocó el surgimiento de diversos proyectos que buscaban soluciones a las problemáticas que planteaba. La dispersión de conocimiento que se produjo fue de tal magnitud que, en 1952, se decidió organizar los primeros congresos de traducción automática para intercambiar experiencias y no duplicar el trabajo.

> Eighteen scholars, including Booth as the only non-American delegate, gathered on the first international conference on Machine Translation at MIT, followed by a meeting later that year in London where some forty linguists met during the International Linguistic Congress. A year later, Machine Translation appeared for the first time in a scholarly textbook written by Andrew and Kathleen Booth. In their book *Automatic Digital Calculators* (Booth and Booth, 1953), aimed at a readership of computer scientists, the authors published a chapter on *Some applications of computing machines* in which Machine Translation was discussed at length. (Vanhoutte, 2013, p. 124)

De estos dos congresos, en tierras americanas y británicas, se llegó a la certeza de que la traducción automática era posible, pero, faltaba la prueba definitiva, que se produjo en 1954.

> A highly publicized demonstration in New York on January 7 of their *701* general-purpose computer, programmed to do translations

> into English of a number of Russian sentences using a total vocabulary of 250 different words. Starting with six rules of syntax formulated by Leon Dostert of Georgetown University, Paul Garvin, also of Georgetown, devised coding which enabled Peter Sheridan of the IBM to program the problem. (Booth & Locke, 1955, p. 8)

IBM, con su primer ordenador comercial, el *701*, había adelantado a Remington Rand y su *UNIVAC I*, con lo que acaparó los titulares. Aunque era un sistema muy rudimentario con un vocabulario muy limitado, mostraba al mundo que los ordenadores eran capaces de traducir de un idioma a otro. Era una cuestión de tiempo que sus posibilidades mejorasen, como se comprobó unos años después con la aplicación *Systran*[37].

En 1956, Peter Toma, de origen húngaro y afincado en California, llevó a cabo una aproximación pragmática al problema de la traducción automática. Estaba convencido de que la solución era que los ordenadores llevasen a cabo un análisis del lenguaje para superar las limitaciones de la traducción palabra por palabra y, basándose en sus ideas, creó *Systran*. La US Air Force la instaló en 1970 para que efectuase la traducción automática de documentación militar del ruso al inglés (Petrits, 2001, p. 4). Según los informes, obtenía una eficacia del 90% y, así, la mayor parte de los documentos no necesitaban modificaciones (Hutchins, 1999, pp. 3-4). Tal fue su éxito que, en 1975, la Comisión Europea compró ciertos derechos sobre este software y creó una versión propia: *EC Systran*. Casi cincuenta años después, lo sigue utilizando[38].

Aunque entre los años 50 y 60 del siglo xx se crearon otros proyectos relacionados con estos que inauguraron las humanidades digitales (Burton, 1981, pp. 4-10), el cambio de paradigma en el tratamiento informático de textos había sido propuesto, en realidad, en 1945 por parte de Vannevar

---

37 *Systran* es un software que aún sigue vivo. Ha ido evolucionando a lo largo del tiempo, incorporando idiomas y refinando su funcionamiento. Su probada fiabilidad hace que se utilice como plataforma de prueba de las últimas novedades algorítmicas, por ejemplo, los *transformers* para el procesamiento del lenguaje natural (Pham *et al.*, 2021, p. 842).

38 En 2010, la Comisión Europea tuvo que pagar 12 millones de euros al grupo Systran por infracción de copyright al considerar que había copiado determinadas características del software original sin permiso. Para más información, véase la Sentencia del Tribunal General (Sala Tercera) de 16 de diciembre de 2010. Caso T-19/07, Systran SA y Systran Luxemburgo SA *vs* Comisión Europea, European Court Reports.

Bush, en su conocido artículo "As We May Think", en el que describía una máquina a la que llamó *Memex*:

> A *Memex* is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. (Bush, 1945, pp. 106-107)

Bush era un gran conocedor de las computadoras. Años antes de inventarse el ordenador electrónico, ya diseñaba dispositivos analógicos, el más famoso de ellos, el analizador diferencial, resolvía ecuaciones diferenciales ordinarias (Bush, 1931, p. 447) y sirvió de inspiración para la construcción del *ENIAC* (Ralston & Reilly, 1983, p. 535). Pero, en este caso, iba un paso más allá. Imaginó una máquina con un sistema de indexación textual que tenía la posibilidad de visualizar los documentos que se buscaban. Sin embargo, lo realmente innovador era su sistema para relacionar elementos.

> It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the Memex. The process of tying two items together is the important thing. (Bush, 1945, p. 107)

El *indexado asociativo*, como lo denominó Bush, era un sistema que permitía enlazar documentos y pasar de uno al otro de forma automática e inmediata, una idea que inspiraría proyectos posteriores que buscaban aplicar la informática al ámbito lingüístico, tal como hizo el filósofo y sociólogo Theodor Holm Nelson en la década de los 60.

> The original idea was to make a file for writers and scientists, much like the personal side of Bush's *Memex*, that would do the things such people need with the richness they would want. (Nelson, 1965, p. 84)

Ante esa necesidad evidente de que el ordenador se adaptase para manejar información más allá del procesado de datos científicos y empresariales, que fuese capaz de trabajar con documentos textuales y crear relaciones

entre ellos, tomando el *Memex* de Bush como inspiración, Nelson propuso en 1965 el *hipertexto*:

> A body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper. It may contain summaries, or maps of its contents and their interrelations; it may contain annotations, additions and footnotes from scholars who have examined it. (Nelson, 1965, p. 96)

Lo hizo en el marco del proyecto *Xanadú*[39], que buscaba optimizar la forma de gestionar los textos informáticos para que el ordenador ayudase a los investigadores en su día a día. No tenían que adaptarse ellos a la máquina, sino la máquina amoldarse a sus necesidades. Crear mapas, resúmenes, anotaciones y enlaces entre documentos debía convertirse en algo natural en un ordenador. Y aunque el concepto era perfecto, su ejecución en ese momento era irrealizable. La interacción humano-ordenador aún era muy rudimentaria y, en este sentido, faltaba por inventarse un elemento hoy en día cotidiano: el ratón.

En paralelo al trabajo de Nelson, Douglas Carl Engelbart, ingeniero eléctrico estadounidense, había creado el *Augmentation Research Center* (ARC), un centro de investigación cuyo objetivo era mejorar la capacidad del ser humano para lidiar con el volumen de información textual generada mundialmente, ya que estaba creciendo de forma exponencial.

> Engelbart's team wanted to invent new methods to process effectively the large volumes of information that knowledge workers use. Peter Drucker had coined the term *knowledge workers*. Drucker defined knowledge as the systematic organization of information. Knowledge workers were people who create and apply knowledge to productive goals. In the late 1960s, Drucker's *The Age of Discontinuity* predicted that a social transformation would occur in the last quarter of the 20th century. The U.S. economy would shift from an economy that manufactured products to one that was based primarily on the exchange of knowledge. (Barnes, 1997, p. 21)

---

39 Nelson creó su proyecto *Xanadú* en 1960 para mejorar la interacción con los ordenadores, pero nunca llegó a ver la luz (Rosenzweig, 2001, p. 548). Este se basaba en el mismo concepto del *hipertexto*, desarrollado años después, que fue, en realidad, el que trascendió.

En los albores de la nueva era de la información, se requerían nuevas técnicas que facilitasen su acceso y gestión. Los ordenadores permitían su almacenamiento, pero una acción tan sencilla como pasar de un documento a otro, con el teclado y con las rudimentarias interfaces textuales, requería efectuar una secuencia de pasos que estaban lejos de ser intuitivos.

Fue en el otoño de 1968 cuando, finalmente, Engelbart y su equipo mostraron públicamente, en una demostración conocida como *The Mother of All Demos,* el fruto de sus investigaciones: el ratón, un dispositivo que iba a cambiar para siempre la forma de interactuar con los ordenadores. Además, este elemento periférico no venía solo, sino que era una parte integrante del *NLS*[40]: una combinación de hardware y software que consistía en una computadora gobernada por un sistema operativo que mostraba una interfaz hombre-máquina evolucionada. Era el engranaje que le faltaba a la idea de Nelson: un ordenador con el que se interactuaba por medio de un sistema de ventanas y manejado por el ratón que, desplazándolo sobre la mesa, permitía saltar rápidamente de un documento a otro, escribir un correo electrónico o tener una videoconferencia con otro miembro del equipo, todo ello de forma intuitiva. Comenzaba una nueva era en el trabajo colaborativo. La oficina sin papel se había inventado[41]. La gestión electrónica de los documentos eliminaba la barrera de la distancia y los equipos de investigadores podrían trabajar en un mismo proyecto independientemente de su localización geográfica. Faltaba universalizar el concepto, extenderlo mundialmente.

Tuvieron que pasar más de 20 años para que, en octubre de 1990, conforme se expandía la conexión a Internet por las universidades y centros de investigación, se produjese esa universalización del acceso a la información. En octubre de 1990, Tim Berners-Lee, físico británico hijo de programadores, que había investigado cómo podía contribuir el hipertexto al acceso e intercambio de información entre investigadores, se puso a desarrollar la que sería la World Wide Web basándose en la idea del hipertexto, ante la ausencia de un sistema comercial que implementase su idea de colaboración documental. En navidades de ese año, el primer navegador de Internet se conectaba a un servidor que se había programado para servir un conjunto de páginas *HTML* (Berners-Lee, 2000, pp. 28-30): había nacido la web. En 1990 había 2 617 586 usuarios de Internet;

---

40 El nombre de *NLS* proviene de *oNLine System*.

41 "This project was the first prototype of a paperless office. No paper changed hands during the communication exchange, and access to information was perceptually immediate" (Barnes, 1997, p. 21).

en 2020, 4 699 887 523, lo que viene a significar que un 60% de la población mundial lo utiliza.

> Estamos aún en la fase del *incunable del hipertexto*, pero la hoja de ruta del cambio ya la están escribiendo los millones de nativos digitales en todo el mundo que se están incorporando a la enseñanza, a la sociedad, al trabajo, y que demandan nuevos modelos de relación con la información y el conocimiento, nuevos modelos textuales que permitan a la segunda textualidad, a la tercera oralidad, desarrollarse en todas sus posibilidades. (Lucía Megías, 2012, p. 140)

En 1951, el padre Busa presentó, en la *XVIII World Conference of Documentation*, un primer volumen de ejemplo de las concordancias de las obras de Santo Tomás de Aquino. Hoy en día, las concordancias contenidas en todos los volúmenes se pueden obtener en el hipertexto que describió Nelson en 1965, con unos simples clics y movimientos del ratón que presentó Engelbart en 1968, a través de la World Wide Web que creó Berners-Lee en 1990. La universalización del acceso a la información ya es un hecho. Habían nacido las humanidades digitales.

2.4. La irrupción de las humanidades digitales en España y su aplicación filológica

Mientras que las computadoras se extendían por las empresas y administraciones americanas, en España hubo que esperar hasta 1958 para que llegase el primer ordenador, un *IBM 650*[42], destinado a la compañía ferroviaria RENFE; poco tiempo después, lo hizo un *UNIVAC UCT*, esta vez para la Junta de Energía Nuclear (Arroyo Galán, 2005, p. 49). Sin embargo, pese a estas rápidas adquisiciones iniciales, no sería hasta diez años más tarde, en 1968, cuando se pondría en funcionamiento la primera máquina dedicada a la investigación en el —expresamente construido para tal fin— Centro de Cálculo de la actual Universidad Complutense de Madrid: un *IBM 4090*. Se trataba de un computador cedido por el fabricante en cuestión, que llegaba a ocupar toda una planta y que ya había estado operativo previamente en el CERN de Ginebra y en la Universidad de Heidelberg (López & Munarriz, 2021, pp. 22-30).

El Centro de Cálculo se consideraba una instalación singular, dado que, ya desde sus inicios, se incentivó el uso de la nueva máquina más allá de los cálculos científicos. Se organizaron seminarios para reunir humanistas con el objetivo de que aportasen ideas innovadoras, como el de Generación Automática de Formas Plásticas, que comenzó a celebrarse durante el curso 1968-1969 con una periodicidad anual (Castaños Alés, 2000, p. 86). Además, se crearon becas en "un empeño en que artistas, arquitectos o músicos aprendieran un lenguaje de programación" (López & Munarriz, 2021, p. 28).

Fue, precisamente, en estas instalaciones donde comenzó su carrera uno de los primeros investigadores que se dedicó a las humanidades digitales: Francisco Marcos Marín. En un artículo fechado en 1971, describía un sistema de traducción automática inglés-francés, que llevaba cinco años desarrollándose en la Universidad de Montreal, y las implicaciones que tendría su adaptación al español (Marcos Marín, 1971, p. 313). Era el mismo año en el que la Real Academia Española (RAE) decidía investigar sobre las posibilidades de aplicación de los ordenadores a sus diccionarios contratando a dos investigadores: Ignacio Soldevila Durante, para estudiar la posible informatización del diccionario académico, y el propio Francisco

42 Actualmente se encuentra expuesto en el Museo Nacional de Ciencia y Tecnología de A Coruña.

Marcos Marín, para colaborar en el *Diccionario Histórico* (2009, p. 392)[43]. Sin embargo, la RAE no disponía de equipamiento informático y, por tanto, también fue el Centro de Cálculo de la Universidad Complutense de Madrid el que acogió, *de facto*, los primeros proyectos de humanidades digitales en España. Allí fue donde se desarrolló el primer analizador sintáctico automático de español (Campo *et al.*, 1973, p. 16) y el primer atlas lingüístico plurilingüe que hacía uso de un ordenador para su representación gráfica (Ariza *et al.*, 1973, p. 12).

La producción científica de Marcos Marín es cuantiosa y su revisión completa supera los objetivos de este trabajo, pero cabe destacar su liderazgo como director científico del equipo que creó *Admyte*, un archivo digital de manuscritos y textos españoles, que incorporaba un buscador que permitía la consulta de las bases de datos, de las transcripciones y de los facsímiles digitales (Marcos Marín, 1994, pp. 196-217). Fue editado en tres CD-ROMs y, actualmente, aún es accesible en línea bajo suscripción[44]:

> El trabajo en equipo se puso de manifiesto en *Admyte 0*, disco que contenía un volumen ingente de información de diversa naturaleza: 64 textos medievales transcritos por diversos investigadores en el marco que les brindaba *el Seminary of Medieval Spanish Studies* de la Universidad de Wisconsin-Madison; los catálogos generales de manuscritos e incunables en español (BETA/BOOST), catalán (BITECA/BOOCT) y portugués (BITAP/BOOPT); un programa de recuperación y análisis textual para los estudios lingüísticos y literarios (TACT); y un programa para preparar ediciones críticas capaz de cotejar hasta 30 versiones diferentes de la misma obra (UNITE).
>
> Los materiales de *Admyte I* fueron seleccionados y elaborados *ex professo* por el equipo editorial, que trabajó en estrecha relación con la Biblioteca Nacional en el marco de la celebración del V Centenario del Descubrimiento de América. En total, se recogieron las transcripciones y las imágenes digitalizadas de 61 obras

43 El objetivo inicial era que Marcos Marín colaborase en los arabismos del *Diccionario Histórico*; sin embargo, su director, Rafael Lapesa, decidió enviarlo a la Escuela de Pisa, que en aquel momento era un referente en lingüística computacional (2009, p. 392).

44 Entre el equipo de *Admyte* estaba Charles B. Faulhaber, que, como se comentará más adelante, será el principal responsable de crear la web *PhiloBiblon* (Faulhaber, 2014, p. 17), un metacatálogo en línea de bibliografías ibéricas.

> transmitidas en libros incunables y postincunables, que dan una idea muy precisa del panorama cultural español y europeo en los años del Descubrimiento de América.
>
> Respecto de estos dos discos previos, *Admyte II* presenta dos cambios de importancia: sólo ofrece las transcripciones de las obras sin imágenes digitalizadas que las acompañen; a cambio, el número de los textos ha crecido notablemente, pues se han incorporado nada menos que 165 nuevas ediciones de obras correspondientes en su mayor parte a la España medieval, aunque tampoco falte un importante testimonio renacentista, el Lazarillo de Tormes en dos de sus testigos principales, y hasta una joya del barroco literario español, el manuscrito Chacón con la obra poética de don Luis de Góngora. Además, *Admyte* hace justicia al romancero por partida doble, pues remedia su ausencia de nuestro catálogo de obras medievales (BETA/BOOST) y corrige el injustificable olvido de este fascinante corpus en los dos discos previos de *Admyte*[45].

Otro de los pioneros españoles en las humanidades digitales fue el lexicógrafo Manuel Alvar Ezquerra, coetáneo de Marcos Marín, quien, a mediados de los años 70, tuvo el privilegio de trabajar junto a Bernard Quemada[46] en el *Trésor de la langue française*, un diccionario histórico en el que se utilizaron medios informáticos para su creación (Marcos Marín, 2009, p. 393). Su primer artículo sobre humanidades digitales se publicó a mediados de los 70 y trata la obtención de índices de rimas y sufijos utilizando una computadora, con el fin de evitar el trabajo manual y, además, acelerando con ello el proceso (Alvar Ezquerra, 1976, pp. 35-36). Pero sus innovaciones metodológicas no quedaron relegadas a la investigación, sino que, además, las llevó a las aulas desde muy temprano, al hacer uso de la automatización de los estudios lingüísticos en sus cursos de Lingüística Aplicada en la Universidad Complutense de Madrid (Pêcheux, 1978, p. 7).

---

45 La cita está extraída de la sección *Historia* de la web *Admyte*, https://www.admyte.com/admyteonline/historia.htm [consulta: 02/04/2023].

46 Bernard Quemada fue un lexicógrafo nacido en San Sebastián, de padre español y madre francesa, que se formó y trabajó en Francia. Estuvo al frente de "l'elaboració del *Trésor de la Langue Française*, un dels primers corpus textuals digitalitzats que es van crear a Europa durant la segona meitat del segle xx" (Cabré Castellví, 2019, pp. 714-715).

Una década después de la instalación del primer computador para uso científico en España, ya era evidente que los ordenadores no solo servían para hacer cálculos, sino que se podían utilizar en las investigaciones humanísticas. Las aportaciones en el campo de la lingüística computacional eran cada vez más numerosas y solo desde estas circunstancias se explica que, en 1983, Montserrat Meya y María Felisa Verdejo creasen la revista científica *Procesamiento de Lenguaje Natural* en la Facultad de Informática de la Universidad del País Vasco, con unas líneas de trabajo abiertas a la colaboración entre distintas ramas de conocimiento, tal como expresaban en su presentación:

> El propósito que perseguimos con ella es, ante todo, crear un órgano especializado de comunicación a través del cual podamos conocer de forma directa los trabajos, investigaciones y proyectos que se realizan en España dentro del área del Proceso del Lenguaje Natural. Por ello, al tratarse de una tarea interdisciplinaria, nos hemos reunido en un mismo empeño profesionales de diferente formación: ingenieros, informáticos y lingüistas. (Meya, 1983, p. 1)

Esta interdisciplinariedad ya venía reflejada en la formación de sus editoras, dado que Meya[47] es lingüista y Verdejo, informática. Meya trabajaba para Siemens, mientras que Verdejo, al igual que Marcos Marín, se inició y ejerció en las humanidades digitales en el marco del Centro de Cálculo de la Universidad Complutense de Madrid:

> El día 30 de octubre de 1975, en la Universidad de París VI (Institut de Programmation) leyó su tesis doctoral de Tercer Ciclo, sobre el tema "Un estudio del lenguaje natural y su aplicación a un diálogo en castellano con un robot" la Analista del Centro de Cálculo de la Universidad Complutense de Madrid Srta. María Felisa Verdejo; la tesis fue juzgada con la calificación de "Très honorable" por un tribunal presidido por el profesor Arsac y entre los miembros del mismo figuraba D. Ernesto García Camarero Director del Centro de Cálculo de la Universidad Complutense. (García Camarero, 1976, p. 102)

47 Montserrat Meya recibió el premio Ada Byron a la Mujer Tecnóloga en el 2014, un galardón creado por la Facultad de Ingeniería de la Universidad de Deusto y que puede consultarse en https://www.deusto.es/es/inicio/somos-deusto/facultades/ingenieria/eventos-y-premios/premio-ada-byron [consulta: 27/08/2023].

El primer artículo de Verdejo (1976, p. 3)[48] sobre humanidades digitales parte de su tesis doctoral: un sistema de interacción de pregunta-respuesta en español con una máquina, dotado de cierta inteligencia en sus respuestas, en el cual no solo aplica el procesamiento de lenguaje natural, sino también técnicas de inteligencia artificial, los dos campos en los que Verdejo ha basado mayormente su investigación. Esta dualidad la llevó a fundar, en 1983, la *Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN) junto a la revista antes mencionada, así como la *Asociación Española para la Inteligencia Artificial* (AEPIA). Como reconocimiento a su trayectoria, en 2014 se le concedió el Premio Nacional de Informática José García Santesmases y en 2016 fue nombrada *Doctora Honoris Causa* por la Universidad de Alicante[49].

Investigadores como Marcos Marín, Alvar, Meya y Verdejo, entre otros muchos, impulsaron el campo lingüístico de las humanidades digitales en los años 70, que se consolidó y se expandió a otras universidades españolas en los 80. Buena muestra de ello es el hecho de que el primer número de *Procesamiento de lenguaje natural* esté formado por artículos de investigadores de las universidades del País Vasco, Castilla-La Mancha[50], Valencia, Barcelona, Granada y Complutense de Madrid, además de otros dos que trabajaban en las empresas privadas IBM y Siemens[51].

---

48 El primer artículo de Verdejo, en realidad, es anterior y en colaboración, sobre la creación de un preprocesador para un lenguaje de programación, el *ALGOL 60* (Garijo & Verdejo, 1973). Un *preprocesador* es una aplicación que, dada una entrada en un lenguaje formal, lleva a cabo un análisis léxico y sintáctico y produce una salida en otro lenguaje que permite su traducción directa, con el uso de un compilador, a código interpretable por el ordenador.

49 Según atestigua el currículum que se elaboró para su nombramiento como *Doctora Honoris Causa* por la Universidad de Alicante y que se puede consultar en https://web.ua.es/es/protocolo/documentos/eventos/honoris/verdejo-maillo-felisa-2016/apunte-biografico-de-felisa-verdejo-maillo.pdf [consulta: 02/04/2023].

50 En aquel momento era el Colegio Universitario de Ciudad Real, que después se ha convertido en la Universidad de Castilla-La Mancha.

51 Los artículos están firmados en este orden por los siguientes investigadores: F. Ares de Blas (Facultad de informática, San Sebastián), N. Antonio Campos (Colegio Universitario, Ciudad Real), F. Casacuberta y E. Vidal (Centro de Informática, Valencia), M. Meya (Siemens, Múnich), J. Rafel (Dpto. Catalán. Universidad de Barcelona), J. Rubio Ayuso y M.C. Carrión Pérez (Universidad de Granada), Martín S. Ruipérez (Facultad de Filología, Universidad Complutense), L. de Sopeña (Centro Científico IBM, Madrid), M.F. Verdejo (Facultad de Informática, San Sebastián).

Fue una década marcada por una actitud aperturista que propició la colaboración en proyectos internacionales como *EUROTRA*, el traductor multilingüe europeo. Había un interés creciente en este tipo de herramientas, por lo que, no solo hubo inversión pública en la traducción automática del español, sino que las grandes empresas informáticas también crearon sus propios proyectos, como IBM con *MENTOR* o SIEMENS con *METAL* (Marcos Marín, 2009, pp. 394-395).

Pero, pese a este interés creciente en la aplicación de la computación en el campo lingüístico, su incorporación en la vertiente literaria de la filología fue más tardía:

> A finales de los años 80 se conocían algunas iniciativas de filólogos que empleaban herramientas computacionales, pero eran rarísimas excepciones, y desde luego ninguna tuvo impulso institucional ni ayudas económicas antes de los años 90. Fue en 1990 cuando el Gobierno Español, con motivo de las Conmemoración del V Centenario del Descubrimiento de América, creó una *Comisión Nacion*al y se nombró director del área de *Industrias de la Lengua de la Sociedad Estatal para la Ejecución de los Programas del Quinto Centenario (1990-1992)* a Francisco Marcos Marín, profesor de la Universidad Autónoma de Madrid, que había estado en contacto con grupos norteamericanos en uno de los proyectos primitivos más ambiciosos que hoy conocemos como *PhiloBiblon*, base de datos biobibliográfica sobre textos romances escritos en la Península Ibérica en la Edad Media y el Renacimiento. (López Poza, 2020, p. 133)

En la década de los 90 es cuando hizo su aparición la World Wide Web y, con ella, la universalización de Internet. Todo el mundo quería estar en ese escaparate mundial que se estaba extendiendo rápidamente más allá del ámbito investigador y, por supuesto, los grandes proyectos humanísticos no fueron ajenos a su influencia[52]. Tal fue el caso de *PhiloBiblon* que, después de su incorporación en *Admyte,* publicaba su versión en línea en 1997. Aunque actualmente está formado por un compendio de diferentes

52 José Manuel Lucía (2010) destaca, en este sentido, tres de ellos en cuanto a la filología hispánica que apostaron por la publicación en línea como el futuro: *LEMIR*, que comenzó a gestarse entre 1994 y 1995; *PhiloBiblon*, en 1997; y la *Biblioteca Virtual Miguel de Cervantes*, que tuvo su primera versión en julio de 1999.

catálogos bibliográficos de textos ibéricos, sus inicios fueron en la década de los 70 y se limitaba a la *Bibliography of Old Spanish Texts* (BOOST):

> Se concibió en un primer momento como un paso inicial necesario para el empleo del ordenador en la confección del Diccionario de español medieval (DOSL) en el que la Universidad de Wisconsin, Madison, viene trabajando desde hace aproximadamente medio siglo. Este ambicioso proyecto lexicográfico tiene como primer propósito la formación y publicación de un amplio corpus léxico por medio de citas, reflejando el empleo de cada término tal como lo presentan los diferentes documentos —impresos o manuscritos— producidos antes del año 1501. (BOOST, 1984, p. XVII)

Estos materiales dieron lugar a la *Bibliografía Española de Textos Antiguos* (BETA), que se recogieron en 1993 en un CD-ROM, como parte de *Admyte*. De igual manera, Beatrice Concheff desarrolló en 1985 la *Bibliography of Old Catalan Texts* (BOOCT), que acabó dando lugar a la *Bibliografia de Textos Antics Catalans, Valencians i Balears* (BITECA)[53]. Ambas bibliografías completaron la perspectiva ibérica con la *Bibliografia de Textos Antigos Galegos e Portugueses* (BITAGAP), derivada, en última instancia, de la *Bibliography of Old Portuguese Texts* (BITAP/BOOPT)[54]. BITECA y BITAP también formaron parte del CD-ROM de *Admyte* antes de llegar a estar disponibles en la red, en este último caso ya como BITAGAP y a diferencia de la *Bibliografía de Poesía Áurea* (BIPA), que se ha sumado a este portal ya avanzado el siglo xxi[55]. BIPA había permanecido hasta entonces

53 BITECA, en un principio, era tan solo la *Bibliografia de Textos Catalans Antics*, aunque se corrigió el nombre a partir de un convenio con la Acadèmia Valenciana de la Llengua.

54 "Deriva de la *Bibliography of Old Spanish Texts* (BOOST), cuya primera edición impresa vio la luz en 1975 (Cárdenas *et al.*, 1975) como repertorio de textos y manuscritos medievales en castellano, una herramienta ancilar del *Dictionary of the Old Spanish Language* de Madison. Con el tiempo BOOST, rebautizada como BETA (*Bibliografía Española de Textos Antiguos*), junto con sus congéneres BITAGAP (*Bibliografia de Textos Antigos Galegos e Portugueses*) y BITECA (*Bibliografia de Textos Antics Catalans, Valencians i Balears*) pasaron al formato CD-ROM como parte de ADMYTE (*Archivo Digital de Manuscritos y Textos Españoles*) en 1993" (Faulhaber, 2009, p. 191).

55 "La *Bibliografía de la Poesía Áurea* es una base de datos digital de la poesía española de los siglos xvi y xvii recogida en fuentes manuscritas e impresas. La versión actual, parcial e incompleta, se ha montado únicamente para hacer pruebas" (*BIPA* 2023). Los responsables de BIPA, según su web el 03/02/2023, son: Ralph DiFranco (University of Denver) y José J. Labrador Herraiz (Emérito, Cleveland State University).

offline y el acceso se llevaba a cabo mediante consultas personales que los investigadores hacían a sus creadores, José J. Labrador y Ralph DiFranco. Aunque la base académica de *PhiloBiblon* se encuentra en la University of California, la mayoría de sus responsables pertenecen a universidades españolas, en las que diferentes investigadores principales han coordinado sucesivos proyectos de investigación de financiación pública, de manera independiente para cada una de estas bibliografías[56].

A pesar de que acabó derivando en la creación de *PhiloBiblon*, la BOOST se originó, en realidad, en el *Hispanic Seminary of Medieval Studies* (HSMS) de la Universidad de Wisconsin-Madison como un proyecto complementario para crear un diccionario electrónico: el *Dictionary of the Old Spanish Text* (DOSL), cuya versión en papel se venía elaborando desde 1935. En 1971, se estudió la posibilidad de generarlo a partir de un corpus electrónico del español medieval y se llegó a la conclusión de que era necesario crear previamente una bibliografía que recogiese los ejemplares de incunables que existían, con el fin de seleccionar el material para confeccionarlo (Nitti, 1978, p. 43). En 1974 apareció la primera versión de BOOST, a la que le siguieron dos más en papel. Por su parte, la creación del corpus electrónico comenzó en 1979 con la edición, en microficha, de *The Concordances and Texts of the Royal Scriptorium Manuscripts of*

56 El equipo de BETA, según su web el 22/07/2024, está formado por Charles B. Faulhaber (University of California, Berkeley), Ángel Gómez Moreno (Universidad Complutense de Madrid), Nicasio Salvador Miguel (Universidad Complutense de Madrid), Antonio Cortijo Ocaña (University of California, Santa Barbara), María Morrás, (Universitat Pompeu Fabra / Oxford University), Óscar Perea Rodríguez (University of San Francisco), Álvaro Bustos Táuler (Universidad Complutense de Madrid), José Luis Gonzalo Sánchez-Molero (Universidad Complutense de Madrid), Almudena Izquierdo Andreu (Universidad de Salamanca) y Patricia García Sánchez-Migallón (École normale supérieure de Lyon). Los responsables de BITECA, según su web el 22/07/2024, son: Gemma Avenoza (†) (Universitat de Barcelona, Institut de Recerca en Cultures Medievals), Lourdes Soriano (Universitat de Barcelona, Institut de Recerca en Cultures Medievals) y Vicenç Beltran (Universitat de Barcelona, Università di Roma "La Sapienza"). Y, finalmente, los recopiladores de BITAGAP desde 1988, según su web el 22/07/2024, son: Arthur L-F. Askins (University of California, Berkeley), Harvey L. Sharrer (University of California, Santa Barbara), Martha E. Schaffer (University of San Francisco) y Aida Fernanda Dias (†) (Universidade de Coimbra). En asociación desde el 2008 con: Cristina Sobral (Faculdade de Letras, Universidade de Lisboa), Pedro Pinto (Centro de Estudos Históricos, Universidade Nova de Lisboa), Filipe Alves Moreira (Universidade Aberta), Mariña Arbor Aldea (Facultade de Filoloxía, Universidade de Santiago de Compostela), Maria de Lurdes Rosa (Departamento de História, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa) y Ricardo Pichel Gotérrez (Universidad de Alcalá).

Alfonso X[57]. Desde el 2011, la colección se publica en la web del seminario como *Biblioteca digital de textos del español antiguo*[58]. Esta serie sirvió de inspiración al *Seminari de Filologia i Informàtica* de la Universidad Autónoma de Barcelona para crear el *Arxiu Informatitzat de Textos Catalans Medievals*, coordinado por Lola Badia, José M. Blecua, Glòria Claveria, Josep Pujol, Amadeu Soberanas y Joan Torruella (1995). El proyecto, en realidad, comenzó en 1986 "amb vista a confeccionar un fons de textos apte per a una sèrie d'ánalisis informàtiques que pretenen acostar-nos amb major precisió a la interpretació i organització de les dades paleogràfiques, textuals i lingüístiques proporcionades pels manuscrits originals" (Torruella & Lawrance, 1988, p. 31). Aunque las primeras microfichas con los textos y las concordancias respectivas generadas con la ayuda del software *Oxford Concordance Program* se publicaron entre 1995 y 1996 (Torruella, 1991, p. 245)[59], ya se habían aplicado técnicas computacionales sobre este corpus de cancioneros catalanes medievales en 1992, para generar un índice de frecuencias de las rimas del *Cançoner L*, así como un rimario alfabético en el que se mostraba, junto a su autor, la consonancia, frecuencia y palabras-rima utilizadas en esa colección poética (Torruella, 1992, pp. 5-9). Este es un ejemplo temprano de la rentabilidad que suponía el tratamiento de corpus desde las humanidades digitales para la obtención de resultados de interés literario y no exclusivamente lingüístico.

También en 1995, y con más de 30 años de diferencia respecto al primer corpus electrónico de la lengua inglesa[60], la RAE acometía el proyecto de crear dos corpus: el *Corpus de referencia del español actual* (CREA) y el

57 Se puede consultar la historia completa del HSMS en http://www.hispanicseminary.org/history-en.htm [consulta: 14/04/2023].

58 A fecha de la consulta, dispone de 12 corpus de distinta temática que agrupan un total de 382 textos de libre acceso a través de http://www.hispanicseminary.org/textconces.htm [consulta: 14/04/2023].

59 *Cançoner L* (Barcelona, Biblioteca de Cataluña, ms. 9); *Espill* de Jaume Roig (Ciudad del Vaticano, Biblioteca Apostóllica, ms. Latín 4806); *Cançoneret de Ripoll* (ms. 129 del fondo de Ripoll del Archivo de la Corona de Aragón); *Cançoner dels Masdovelles* (Barcelona, Biblioteca de Cataluña, ms. 11) y *Cançoner del Marquès de Barberà* (Biblioteca del Monestir de Montserrat, ms. 992). Se trataba de cinco juegos de microfichas, que, posteriormente, se completaron con otros cinco: el *Cançoner Vega-Aguiló* (Barcelona, Bblioteca de Catalunya, ms. 7 i 8); *Cançoner de París* (París, Biblioteca Nacional, ms. esp. 225); *Cançoner de l'Ateneu* (Barcelona, Biblioteca de l'Ateneu Barcelonès, ms. 1); *Jardinet d'orats* (Barcelona, Biblioteca de la Universitat, ms. 151); y *Cançoner de Saragossa* (Zaragoza, Biblioteca Universitaria, ms. 184).

60 En 1964 se publicaba el primer corpus electrónico del inglés, el *Brown University Standard Corpus of Present-Day American English*, más conocido como *Brown Corpus*

Corpus diacrónico del español (CORDE), que, respectivamente, atendían a la perspectiva sincrónica y diacrónica de la lengua, cuya primera versión se ponía en línea en 1998 (Rojo, 2016, p. 200).

> Ambos conjuntos son complementarios, de manera que el CREA contiene los textos pertenecientes a los últimos treinta años de historia del español, mientras que el CORDE se ocupa de todo lo demás. El carácter integrado de ambos corpus se refleja en la previsión de que los textos pertenecientes a periodos que, por el paso del tiempo, vayan quedando fuera del ámbito del CREA pasarán a formar parte de CORDE. (Sánchez Sánchez & Domínguez Cintas, 2007, p. 137)[61]

No obstante, en el diseño del CREA se estableció un porcentaje de textos distinto a cada quinquenio, comenzando por un 10% y añadiendo un 5% adicional en cada salto temporal. De esta manera, los que se situaban entre el 1975 y 1979 suponían un 10%, los comprendidos entre 1980 y 1984, un 15%, y así sucesivamente. Esta característica implicaba que, al traspasarlos al CORDE, estos porcentajes debían reajustarse para que todos tuviesen el mismo peso, lo que habría requerido, en última instancia, la eliminación de textos[62]. Es por ello que, finalmente, se decidió no trasvasar información al CORDE, cerrar la incorporación de nuevos regis-

---

(Francis & Kucera, 1979). La primera versión estaba basada en los textos de los libros y periódicos publicados durante 1961 (Francis, 1965, p. 267). Respecto a los primeros corpus de español anteriores al CREA y CORDE, Rojo (2015) enumera los siguientes: el *Corpus de Lovaina* bajo la dirección de Josse de Kock; ENTREVIS90 y ENTREVIS95, construidos por Kjær Jensen; el *Corpus Vox-Biblograf* (CVB), dirigido por Manuel Alvar Ezquerra; el corpus CUMBRE, dirigido por Aquilino Sánchez; el *Corpus del español mexicano contemporáneo* (CEMC), que ha permitido la creación de varios diccionarios dirigidos por Luis Fernando Lara; los corpus multilingüe CRATER, NERC y PAROLE, surgidos en el marco de proyectos europeos; el *Corpus lingüístico de referencia de la lengua española en Argentina*, *el Corpus lingüístico de referencia de la lengua española en Chile* y el CORLEC, los tres dirigidos por Francisco Marcos Marín; el corpus LEXESP, elaborado por un equipo de lingüistas y psicólogos; *el Corpus of Contemporary Spanish* elaborado por Barry Ife; y ADMYTE, confeccionado por Francisco Marcos Marín, Charles Faulhaber, Ángel Gómez Moreno y Antonio Cortijo Ocaña.

61 Rojo (2015, p. 200) establece el marco temporal del CREA en 25 años.

62 Se puede encontrar una explicación más detallada de los porcentajes de cada periodo temporal y la problemática que acarreaba el pasar los textos del CREA al CORDE en Rojo, 2016.

tros en el CREA en 2008 y ceder el testigo a un nuevo corpus: el *Corpus del español del siglo xxi* (CORPES XXI) (Rojo, 2016, pp. 200-202).

Los proyectos literarios de esta época se apoyaron en la informática para crear bases de datos y eliminar el papel. La difusión de los resultados de las investigaciones en Internet suponía modificar el modelo tradicional de publicación impresa. Era un cambio de paradigma que, a pesar de que tres décadas después se ha consolidado totalmente en el caso de las revistas científicas, en ese momento suponía toda una revolución y un cambio profundo del proceso editorial. En el campo de la filología hispánica, solo unos visionarios se atrevieron a romper con ese modelo tradicional y dar el salto al mundo digital. Tal fue el caso de José Luis Canet, de la Universitat de València, y Sagrario López Poza, de la Universidade da Coruña.

Canet se encargó de dar vida a *LEMIR* a mediados de los 90, pero su relación con la informática había comenzado una década antes, desde la propia aparición de los ordenadores personales. Su interés empezó por las bases de datos relacionales, con las que creó un gestor bibliográfico, hecho que le valió para que Apple lo incluyese entre sus desarrolladores en 1987 y que, en 1990, fuese nombrado director de las Bibliotecas de la Universitat de València con la misión de modernizarlas tecnológicamente. Fruto de ese trabajo, apareció en línea en 1993 la web del *Servicio de Información Bibliográfica* (SIB) de la Universitat de València, experiencia que aprovechó para lanzar posteriormente el portal *LEMIR*[63], enfocado a la literatura medieval y renacentista, en cuyo marco apostó por una de las primeras revistas de literatura con formato únicamente electrónico: *LEMIR. Revista Española Medieval y del Renacimiento* (Canet, 2014, pp. 11-12). Su primer número apareció en 1995-1996 con tres artículos y cinco ediciones de texto. El portal, en su primera versión, además de la revista electrónica, ofrecía noticias y eventos de interés, ediciones de textos de difícil publicación en editoriales comerciales y un par de secciones con entidad propia: *Exemplaria* y *Tirant*, que, posteriormente, se convirtieron en revistas electrónicas. La primera estaba dedicada a la literatura sapiencial y dirigida por Marta Haro[64], mientras que *Tirant*, a cargo de Rafael Beltrán, se centraba en la literatura caballeresca. El primer número de *Exemplaria*, de 1996, recoge un repertorio de novedades bibliográficas, un apartado de textos con dos

63 El portal *LEMIR* fue una evolución de un *gopher* —un servicio de Internet textual que funciona mediante menús jerárquicos— que había montado previamente y que alojaba textos medievales españoles (Canet, 2014, p. 12). Se puede consultar una de las primeras versiones del portal en https://web.archive.org/web/19970607204833/http://www.uv.es/~lemir/ [consulta: 06/04/2023].

64 Este *Boletín de Literatura Sapiencial* recibe el nombre de *Exemplaria* en su número 0, pero ya en el 1 pasa a llamarse *Memorabilia*, tal y como se le conoce hasta hoy.

transcripciones y un último apartado de miscelánea de investigación que da cabida a noticias en general. Por su parte, el primer número de *Tirant* se materializó dos años después con cuatro artículos, seis reseñas, un apartado de bibliografía, otro de novedades bibliográficas y el resumen de una tesis.

El concepto de *LEMIR* fue un éxito, por lo que, en 1998, vio la luz un proyecto más ambicioso, un portal de portales: el *Servidor Web de Literatura Española* "*Parnaseo"* (Canet, 2014, p. 13). Además de *LEMIR*, que albergaba entonces, como hemos visto, las secciones de *Memorabilia* y *Tirant*, se alojó en *Parnaseo* un nuevo portal: *Ars Theatrica*, dirigido por el propio Canet junto a Evangelina Rodríguez y Josep Lluís Sirera, y centrado en el teatro español. En paralelo a estas dos grandes secciones, *Parnaseo* se planteó mantener charlas en tiempo real a través de la propuesta de *Parnachat*[65], así como ediciones facsimilares, enlaces de interés e, inicialmente, dos bases de datos de producción propia con una estrecha relación: *Imprenta valenciana siglo* XVI y *Tipobibliografía valenciana siglos XV y XVI*[66]. La primera de ellas aloja un repertorio abreviado bibliográfico de la producción impresa en Valencia durante el siglo XVI[67], mientras que la segunda[68] lo desarrolla para ofrecer la catalogación tipográfica completa de la imprenta valenciana de los siglos XV y XVI[69].

---

65 A mediados de los 90, el chat era un sistema de comunicación textual en tiempo real muy utilizado. Se basaba en un protocolo de comunicación, el IRC, que permitía establecer comunicaciones privadas y grupos de discusión —salas—. Con el tiempo, surgieron otras alternativas, como el Microsoft Messenger, que también han sido relegadas por las actuales: WhatsApp y Telegram.

66 Se puede consultar la versión de *Parnaseo* del 5 de diciembre de 1998 en https://web.archive.org/web/19981205234850/https://parnaseo.uv.es/ [consulta: 01/04/2023].

67 La base de datos *Imprenta valenciana siglo XVI*, entre otras cosas, ha permitido documentar el proceso de retroceso del valenciano respecto al castellano en la producción literaria de la época y que los textos valencianos publicados eran mayormente de poesía religiosa y profana (Canet, 2004, pp. 20-21).

68 Aunque se puso en línea posteriormente, se empezó a desarrollar a principios de los 90. El proyecto figura activo desde 2010 bajo la dirección de Marta Haro y con José Luis Canet como responsable de la sección (Canet, 2019, p. 455).

69 "Los materiales utilizados en esta base de datos corresponden a tres tesis doctorales realizadas por Purificación Hernández Royo, *La imprenta valenciana de la familia Mey-Huete en el siglo XVI: Producción y Tipografía*, Universitat de València, Facultat de Filologia, octubre de 1994; Gloria Irún de Sojo, *Catálogo gráfico-descriptivo de la Imprenta del Molino de la Rovella: s. XVI*, Universitat de València, Facultat de Filologia, enero de 1995 y Diego A. Romero Lucas, *Catálogo gráfico descriptivo de la Imprenta en Valencia (1473-1530).* Universitat de València, Facultat de Filologia, 16 de febrero

En un momento en el que los facsímiles digitales no existían y las bibliotecas estaban en pleno proceso de informatización de su catálogo completo, la recolección de los datos para la confección de la *Tipobibliografía valenciana siglos XV y XVI* fue un trabajo complejo (Hernández Royo, 1994, pp. 43-44). Desde entonces, esta base de datos ha estado en continuo crecimiento y actualmente aloja más de 900 fichas. Dispone de una intuitiva interfaz organizada en tres niveles de detalle: el buscador, el resultado del filtrado y el contenido propio de cada ficha. Esta estudiada disposición de elementos en una única página agiliza el acceso a cada ficha individual cuando se tienen que consultar varias de ellas, al no tener que volver a una página anterior y recargarla para obtener de nuevo el listado. Además, el contenido de cada una de las fichas es hipermedial, está enriquecido con imágenes que acompañan las descripciones textuales: "cada ficha permite la consulta y ampliación de las imágenes de las portadas, preliminares, colofones, grabados, letras capitales, escudos, filigranas y toda la información editorial pertinente" (Canet, 2019, p. 456).

Desde su aparición, *Parnaseo* no ha dejado de crecer con la creación de nuevas bases de datos, como la de *Carteles teatrales valencianos del siglo XIX*, dirigida por Canet en 2007[70], así como con la incorporación de otras revistas electrónicas, como la histórica *Celestinesca*[71], o la creación de nuevos portales, entre los que destacan los más recientes coordinados por Marta Haro Cortés: *Aula Medieval* y *Portal Celestinesco*.

Beltrán, Canet y Haro, en un momento en el que la publicación en papel era la norma, impulsaron la creación de revistas digitales y bases de datos que, más de veinte años después, siguen publicándose y renovándose. Fueron unos visionarios que supieron aprovechar las posibilidades que brindaba la red como medio de difusión. El tiempo les ha dado la razón. *Parnaseo* ha continuado creciendo y se ha convertido en un servidor web de referencia en el mundo filológico hispánico en el que, actualmente, se alojan diversos portales temáticos, bases de datos, publicaciones electrónicas y su propia línea editorial.

Por su parte, Sagrario López Poza vio en las humanidades digitales la oportunidad para acceder con mayor facilidad a datos sobre los que fundamentar sus investigaciones, así como un mecanismo para difundir

---

de 2005; todas ellas dirigidas por el prof. José Luis Canet. También se incluyen otros materiales recogidos posteriormente"; recuperado de https://parnaseo.uv.es/imprenta/publicacion/presentacion.html [consulta: 01/04/2023].

70 Para otras bases de datos, véase https://parnaseo.uv.es/Bases.htm

71 Bajo la dirección de Joseph T. Snow, que se venía editando en papel desde 1977 y en 2003 pasó a formar parte también de *Parnaseo*.

los resultados que generaba[72]. Por ello, creó el *Seminario interdisciplinar para el estudio de la Literatura Áurea Española* (SIELAE):

> La idea de crear el SIELAE (Seminario interdisciplinar para el estudio de la Literatura Áurea Española) se gestó en 1992 (y se concretó en 1996) al percatarnos de que el estudio de la literatura y la cultura de los siglos XVI y XVII, para una comprensión cabal, precisaba de enfoques no limitados a una disciplina o área de conocimiento. Había que deshacer las fronteras postizas de competencias de áreas impuestas durante los últimos siglos (y que no existían en el Renacimiento y el Barroco) para acometer de manera más integral el estudio desde perspectivas multidisciplinares (López Poza & Saavedra Places, 2014, pp. 285-286).

La coexistencia de filólogos e informáticos en un mismo equipo de trabajo permitió abordar metodologías para el estudio filológico que habrían sido inconcebibles de otra manera, avaladas por la propia circunstancia de que las convocatorias estatales de proyectos incentivaban que se estableciese este tipo de colaboraciones. La primera línea de investigación que demostró los beneficios de un grupo multidisciplinar de tales características fue la relativa a las *Relaciones de sucesos*:

> Corría el año 1993 cuando comenzamos a trabajar en esta línea de investigación y a pensar en el diseño conceptual de una base de datos relacional para catalogar estos impresos, tan numerosos y, paradójicamente, tan desatendidos por los estudiosos hasta entonces. El equipo de investigación para el estudio de las *Relaciones de sucesos* (siglos XVI-XVIII), dirigido por la profesora Sagrario López Poza se puso en contacto con la profesora de bases de datos de la Facultad de Informática de la UDC Nieves R. Brisaboa e inició una colaboración para elaborar el diseño conceptual de la base de datos que dio origen al catálogo (Pena Sueiro & Álvarez García, 2014, p. 338).

---

72 "Sin abandonar los métodos analíticos, críticos o especulativos que caracterizan al área de conocimiento de la Filología Hispánica, queríamos respaldar con datos nuestros estudios y a la vez ayudar a otros en la búsqueda, estudio y difusión de algunos de los campos menos trabajados hasta entonces en la Literatura del Siglo de Oro. Para ello, las nuevas tecnologías fueron la respuesta" (López Poza & Saavedra Places, 2014, p. 286).

López Poza confió a Nieves Rodríguez Brisaboa[73] la parte tecnológica de sus investigaciones. Ambas formaron un equipo en 1994 (Rodríguez Brisaboa *et al.*, 2019, p. 33) que perduró en el tiempo y que, en última instancia, fue el responsable de la aparición de un portal de referencia en el mundo filológico tan relevante como la *Biblioteca Digital Siglo de Oro* (BIDISO).

La primera incursión en la red de López Poza comenzó en septiembre de 1996 con la publicación de la web de *Literatura Emblemática Hispánica*[74]. Su objetivo era alojar la producción del equipo internacional que había reunido y que estaba compuesto por investigadores de las ramas de la computación y de la filología[75]. Esta primera versión ofrecía información sobre sus líneas de trabajo y proyectos, bibliografía específica, noticias relevantes y enlaces a otras páginas académicas de interés. Justo un mes después, junto con Nieves Pena Sueiro, ponía en línea otra página: el boletín de *Relaciones de sucesos españolas en la edad moderna*[76], cuyo primer número recopilatorio se publicaba en enero de 1997[77]. "El grupo se ocupó del mantenimiento y actualización de estas páginas durante 17 años, desde 1996 hasta junio de 2013; fue en este año en el que ambas páginas web se integraron en el portal BIDISO constituyendo una de las cuatro secciones" (Pena Sueiro & Álvarez García, 2014, p. 339). En paralelo a la creación de estas páginas, el equipo estaba embarcado en recopilar y digitalizar ejemplares de relaciones de sucesos dispersos a lo largo de la geografía española para almacenarlos en una base de datos, un trabajo que suponía todo un reto tanto por la cantidad de material que debían manejar como por el

73 Esta colaboración permitió a Rodríguez Brisaboa la creación del *Laboratorio de Bases de Datos*, "cuya línea de investigación aplicada fundamental fue, como no, la de las Bibliotecas Digitales" (Rodríguez Brisaboa *et al.*, 2019, p. 35).

74 Se puede consultar la versión de 1998 en la que se lee, a pie de página, que está en línea desde septiembre de 1996, en https://web.archive.org/web/19981202185628/http://rosalia.dc.fi.udc.es/emblematica/ [consulta: 04/04/2023].

75 Se pueden consultar los miembros del grupo *Emblemática* en la siguiente dirección: https://web.archive.org/web/19991022031636/http://rosalia.dc.fi.udc.es/emblematica/Presentacion.html [consulta: 04/04/2023].

76 El boletín, más tarde, pasó a llamarse BORESU. Aunque la primera versión capturada por Internet Archive es de febrero de 1999, en la portada se indica que está en línea desde octubre de 1996: https://web.archive.org/web/19990219220356/http://rosalia.dc.fi.udc.es/BORESU/ [consulta: 04/04/2023].

77 La sección *Relaciones de sucesos* de BIDISO alberga el repositorio de los boletines publicados desde el número 1.

medio físico en el que se encontraba[78]. Todo el contenido recopilado se pudo consultar públicamente en Internet a partir de 2001 (Pena Sueiro, 2017, pp. 76-79).

López Poza, con la creación del SIELAE, apostó por los equipos multidisciplinares como una nueva forma de abordar las investigaciones humanísticas que, hasta ese momento, habían estado confinadas únicamente al ámbito filológico. Ese nuevo enfoque les permitió avanzar en paralelo en sus investigaciones e intercambiar conocimiento que ha beneficiado tanto a las ciencias humanas como a las ciencias computacionales. Marcó un camino a seguir que, aunque este modelo de interacción y/o cooperación aún no se ha consolidado, cada vez es más habitual.

El uso de Internet como plataforma de difusión de conocimiento también permitió ampliar el concepto clásico de fondo bibliográfico con la aparición de las *bibliotecas digitales* "a texto completo, o mejor dicho a imagen completa de los libros, diarios, mapas, grabados, etc., capaz de ser visualizados mediante cualquier navegador" (Canet, 2000, p. 71), un concepto que implica una evolución respecto a la consulta del catálogo vía web. En este caso, además de la ficha bibliográfica, se proporciona acceso a una edición digital de la obra, siguiendo el modelo iniciado por el *Proyecto Gutenberg*[79]. Algunas de las primeras bibliotecas españolas que se aventuraron a ofrecer este tipo de acceso fueron la *Biblioteca Digital de Catalunya* (BDC), la *Biblioteca Digital de la Universitat de València* (BDUV) y la *Biblioteca Virtual Miguel de Cervantes* (BVMC).

La BDC fue impulsada por el *Consorci de Biblioteques Universitàries de Catalunya* (CBUC)[80] con la finalidad de "proporcionar un conjunto nuclear de información electrónica interdisciplinaria para la totalidad de la comunidad universitaria e investigadora de Catalunya independientemente de donde estas personas ejerzan su actividad" (Anglada & Comellas, 2000, p. 243). El proyecto comenzó a gestarse en 1997 y en 1999 se

78 Se comenzó el proceso con la digitalización de microfilms (Pena Sueiro, 2017, p. 76).

79 "En juillet 1971, Michael Hart crée le Projet Gutenberg pour diffuser gratuitement sous forme électronique les oeuvres littéraires du domaine public. Un projet longtemps considéré par ses détracteurs comme impossible à grande échelle. Site pionnier à tous égards, le Projet Gutenberg est à la fois le premier site d'information sur un réseau encore embryonnaire et la première bibliothèque numérique. Michael numérise lui-même les cent premiers livres" (Lebert, 2005, p. 3).

80 Está formado por las bibliotecas de las universidades públicas catalanas y la Biblioteca de Catalunya.

publicaba su primera versión[81]. A diferencia del *Proyecto Gutenberg*, que tenía como fin ofrecer versiones digitales de material impreso, su objetivo inicial era proporcionar únicamente acceso a diversas bases de datos y revistas electrónicas con la intención de abaratar costes con la suscripción conjunta, aunque "durant el període que s'inicia el 2005 la BDC ha entrat de ple en l'àmbit de les compres/subscripcions a llibres electrònics" (Anglada & Comellas, 2010, p. 10). Actualmente, aloja más de 53 000 recursos electrónicos, con la posibilidad de buscar en su amplio catálogo tanto desde su web como desde los buscadores de las instituciones participantes.

En ese momento en el que nacían las bibliotecas digitales, Canet ocupaba el cargo de *Director del Servei d'Informació Bibliogràfica de la Universitat de València*. Como defensor de la difusión de la información en Internet, en 1999 y con motivo de la celebración del V Centenario de su universidad, propuso la creación del proyecto *Biblioteca Digital*. Su objetivo era poner en la red materiales digitalizados para fines de investigación y "mostrar los fondos valiosos depositados en su Biblioteca Histórica, enlazando así con otro de los proyectos aprobados, *Thesaurus*, encargado de la catalogación y difusión del patrimonio de la Universitat" (Canet, 2000, p. 72)[82]. Al año siguiente, se ponía en marcha el proyecto de digitalización del fondo antiguo a texto completo denominado *Somni* (sueño). Este término deriva del título del incunable *El somni de Johan Johan* de Jaime Gazull, ejemplar único en el mundo, impreso en Valencia, en 1497 y que se conserva en la *Biblioteca Històrica de la Universitat de València* (Millás Mascarós & Escriche Soriano, 2017, p. 4).

La *colección Somni* se integró en el 2008 en el *Repositori d'Objectes Digitals per a l'Ensenyament, la Recerca i la Cultura* (RODERIC), de tipo *Open Access*, y en el 2010 formó parte de *Europeana Regia*[83]. Dado que las primeras digitalizaciones se habían llevado a cabo a partir de microfilms,

---

81 La primera versión de la BDC daba acceso a "BDA Aranzadi (sobre derecho español), Business Source Elite, Econlit, ERIC, Inside, Mathscinet, Medline y The Serials Directory. También se cuenta con la suscripción a las revistas de Academic Press (IDEAL). Además, se da acceso a diversas bases de datos gratuitas catalanas, producidas por universidades y centros de investigación" (Anglada & Comellas, 2000, p. 242).

82 Los materiales que se querían almacenar inicialmente eran prensa periódica antigua (s. XVIII y XIX), libros del s. XV, XVI, XVII, XVIII y XIX, manuscritos, carteles/grabados/mapas y la colección de monedas de la Universitat de València (Canet, 2000, p. 73).

83 *Europeana Regi*a fue un proyecto europeo para digitalizar el fondo bibliográfico de manuscritos de la Edad Media y el Renacimiento, en concreto, la Biblioteca Carolingia, la Biblioteca de Carlos V y la Biblioteca de los Reyes de Aragón y Nápoles.

con motivo de la participación en este proyecto, se realizó de nuevo una captura, pero esta vez directamente de las obras, con el fin de disponer de imágenes de mayor resolución (Millás Mascarós & Escriche Soriano, 2017, p. 5). Actualmente, la colección cuenta con más de 8 000 reproducciones digitales, entre las que se encuentran colecciones de carteles, libretos de fallas, manuscritos, mapas, publicaciones periódicas y libros desde el siglo xv al xx.

Por su parte, la *Biblioteca Virtual Miguel de Cervantes* (BVMC) aparecía en Internet en 1999 como un espacio puramente virtual, es decir, no estaba unido a una biblioteca física. Era un proyecto impulsado por la Universitat d'Alacant y patrocinado por el Banco de Santander y la Fundación Botín (Rovira Soler & Rovira-Collado, 2019, p. 54). Al igual que el resto de bibliotecas digitales, abrió la posibilidad de acceder a las obras de forma totalmente gratuita a través de Internet, hecho que llevó a tratar con especial atención los derechos de autor (Rovira Soler, 2001, p. 68). Desde sus inicios, la BVMC ha sido un portal que ha aportado ideas innovadoras a las humanidades digitales, como la transcripción de las fuentes digitalizadas o la creación de versiones sonoras de las obras. Además, dispone de una sección específica para apoyar la creación de nuevas formas de uso de los datos digitales y del que han salido herramientas de suma utilidad como, por ejemplo, el analizador de versos, que, a partir de un poema, nos indica de forma automática su cadencia, las sinalefas, los acentos y el número de sílabas. Todos sus datos son accesibles en abierto y se ha ido adaptando a los estándares según ha ido pasando el tiempo. Actualmente, tiene la base de datos en el estándar *RDF* y el vocabulario *RDA*[84] (Candela *et al.*, 2018), que se pueden consultar a través de un punto de acceso *SPARQL*[85]. La BVMC, en cualquier caso, no solo se limita a proporcionar acceso abierto a sus datos, sino que también enriquece su contenido a partir de repositorios externos, como es el caso de las fichas de autor, que obtienen parte de su información de *Wikidata*[86]. Es toda una demostración de las posibilidades de la web semántica[87] y del trabajo colaborativo para la creación de conocimiento.

---

84 *RDF* y *RDA* son estándares que facilitan el intercambio de información entre ordenadores definiendo el formato de los datos.

85 *SPARQL* es un lenguaje informático para realizar consultas sobre datos definidos en RDF.

86 *Wikidata* (https://www.wikidata.org/) es una base de datos que permite la edición en abierto de forma colaborativa. Está pensada para utilizarla desde otras aplicaciones, a diferencia de la *Wikipedia*, que presenta la información para lectores humanos.

87 Para un resumen de la historia y el uso de la web semántica véase Hitzler, 2021.

En definitiva, los proyectos filológicos que, en el último cuarto del siglo pasado, utilizaron la informática y se adentraron en la red de redes, sentaron las bases sobre las que se han erigido las investigaciones actuales, cuya presencia en el universo digital ya resulta indispensable. Ciertos humanistas e informáticos comenzaron el camino de las humanidades digitales en España, abrazando la tecnología y la colaboración interdisciplinar como la norma y no la excepción. Impulsaron la digitalización y la difusión pública de contenidos. Su visión de futuro ha permitido que, en pleno siglo XXI, tengamos revistas científicas digitales en abierto, repositorios que se puedan consultar sin suscripción y facsímiles digitales accesibles a golpe de clic. Las humanidades son digitales gracias a ellos.

2.5. Cancionero e imprenta en la red

Los ordenadores han evolucionado, como también lo han hecho sus aplicaciones. En poco más de 50 años se ha conseguido disponer de *smartphones*, unos dispositivos a batería y de bolsillo con más potencia que la computadora utilizada para descifrar códigos en la II Guerra Mundial, que ocupaba una habitación entera y necesitaba una línea de alimentación especial por la cantidad de electricidad que requería para su funcionamiento. Este grado de evolución ha afectado y se ha visto reflejado también en su aplicación a los proyectos humanísticos.

En la actualidad, todos ellos se apoyan en la informática en mayor o menor medida y, de hecho, resulta complejo establecer cuándo hablamos de humanidades digitales y cuándo no; por ello, la asociación *Humanidades Digitales Hispánicas* (HDH)[88] publicó en 2021 un *Documento de recomendaciones para la evaluación y reconocimiento de la investigación llevada a cabo en el ámbito de las humanidades digitales*, en el que propuso ampliar el concepto clásico de resultados de la investigación con una clasificación de los tipos de proyectos que se pueden englobar en el campo de las humanidades digitales[89]. Entre sus propuestas de categorización figuran bases de datos, repositorios y recursos digitales, corpus lingüísticos, herramientas de tratamiento de datos en cualquiera de sus fases, ediciones críticas digitales, así como aplicaciones orientadas a la sociedad y obras audiovisuales, hipermediales[90] y transmediales[91].

Aunque la propuesta de la HDH es reciente, las humanidades digitales no son algo nuevo, como demuestra la gran cantidad de proyectos que llevan años en funcionamiento. Es por esta razón que una revisión completa de todos ellos sería inabarcable como objeto de esta monografía, por lo que, dados nuestros intereses científicos, nos limitaremos a aquellos que tienen la poesía de cancionero y de romancero como objetivo último, así como a otros que estudian sus fuentes, con especial atención a las impresas. Se pretende ofrecer un panorama tan amplio como sea posible al respecto y,

88 La HDH se creó en el 2011 con el fin de promover el desarrollo de las humanidades digitales de habla hispana: https://humanidadesdigitaleshispanicas.es/la-asociacion/organizacion/ [consulta: 30/04/2023].

89 Se puede consultar el documento en https://humanidadesdigitaleshispanicas.es/informes/ [consulta: 30/04/2023].

90 Las obras hipermediales están compuestas por distintos medios, como texto, imágenes, videos y sonidos, que se unen mediante enlaces.

91 Las obras transmediales son aquellas que se despliegan a través de distintos medios de comunicación.

en cualquier caso, que ilustre las tendencias y prácticas de las humanidades digitales en la creación de almacenes de información y bases de datos, así como en ediciones de textos innovadoras.

Si nos centramos en este campo de estudio, *An Electronic corpus of 15th century Castillian "Cancionero" manuscripts*, de la University of Liverpool, es uno de los más importantes proyectos de humanidades digitales en este ámbito. Dorothy Severin lo concibió como una conversión al mundo digital de los siete volúmenes de *El cancionero castellano del siglo xv* de Brian Dutton (1990-1991), de quien tomaba el sistema de identificación de fuentes y poemas. Aunque a día de hoy no está accesible en línea[92], alojaba gran parte del corpus poético cancioneril castellano del siglo xv. El punto de partida eran las detalladas descripciones de manuscritos, en su mayoría a cargo de Manuel Moreno y Fiona Maguire, fundamentalmente, así como las transcripciones de buena parte de los testimonios, que permitían la colación de estos y el establecimiento de variantes de manera automatizada, con lo que ello implicaba para la filología estricta. De manera complementaria, ofrecía imágenes digitales de algunos manuscritos, que permitían la comprobación de las lecciones originales o de algún problema evidente de transmisión textual.

A este mismo ámbito pertenece el proyecto en línea de *Cancioneros Impresos y Manuscritos* (CIM), un portal creado en 2011 por el grupo de investigación del mismo nombre dirigido por Josep Lluís Martos y con sede en la Universitat d'Alacant, cuya nueva versión *responsive*[93], basada en el gestor de contenidos *Drupal* y estrenada a finales del 2022, incorpora diversas herramientas, una de las cuales es *Poesía, Ecdótica e Imprenta* (POECIM), también coordinada por Martos:

> un catálogo en *open access* que nace en la Universidad de Alicante en el marco del grupo internacional CIM, como resultado de sus proyectos de investigación, con publicación periódica anual. Su objetivo principal es la publicación de trabajos científicos para el estudio de las fuentes poéticas impresas, de cancionero y romancero,

92 La URL donde se alojaba originalmente (http://cancionerovirtual.liv.ac.uk) devuelve un mensaje de que no está accesible. Se puede acceder parcialmente gracias a la *Wayback Machine* del proyecto *Internet Archive* en https://web.archive.org/web/20200711215838/http://cancionerovirtual.liv.ac.uk [consulta: 10/02/2023].

93 Una *web responsive* es aquella que adapta el contenido al dispositivo de visualización, por lo que permite su consulta tanto desde un ordenador como desde una tableta o un móvil, variando su diseño para que sea legible, y aprovechando las características del dispositivo en concreto.

> desde el periodo incunable hasta mediados del siglo xvi, para generar, así, un espacio científico que caracteriza y singulariza a este grupo de investigación. Los trabajos publicados en este catálogo representan aproximaciones monográficas a cada una de estas fuentes poéticas, desde una perspectiva material, interna, socioliteraria y ecdótica, como sus principales ejes de estudio, que atienden a la edición impresa y a los ejemplares conservados o conocidos, de los que se ofrece enlace a su ficha en el catálogo del fondo que los conserva y, cuando la hay, también a su digitalización en línea. De cada una de estas fuentes y sus textos, se aporta bibliografía específica y actualizada. (Martos, 2022a)

Esta herramienta se centra en las fuentes impresas de cancionero y las desarrolla, de momento limitándolas al siglo xv[94], como complemento a la atención prestada por el portal de Severin a los cancioneros manuscritos, si bien también hay en CIM una sección dedicada a estos, como veremos. Cada una de las fichas sobre las fuentes incluye información valiosa sobre las ediciones antiguas y sus ejemplares conservados o conocidos, desde una perspectiva material, interna, ecdótica, lingüística y socioliteraria, fundamentalmente. Pero no es solo rica en contenidos, sino también en usabilidad, ya que dispone de un buscador que permite una rápida localización de las fichas que nos interesan ayudándonos de los campos de filtrado habilitados para tal fin.

Asimismo, con una funcionalidad similar, está la sección de *Repertorios abreviados*, que, de momento, ofrece en abierto el *Repertorio Abreviado de Fuentes Impresas del Romancero (1501-1552)* (RAR16), desarrollado por Mario Garvin (2022), y las *Fuentes del Romancero*, recopiladas por Virginie Dumanoir (2023). Además de estos, hay en proyecto la elaboración de otros cinco repertorios abreviados, que suponen una catalogación completa de fuentes poéticas de la poesía impresa incunable y post-incunable en castellano y en catalán, así como de todo el romancero manuscrito e impreso desde sus orígenes hasta 1552[95]:

> A partir de las necesidades detectadas, si no generadas, por nuestras propias líneas de estudio, se dará lugar a siete repertorios, organizados en tres grupos. Ofreceremos tres repertorios abreviados

94 Además de una sección sobre algunas fuentes impresas de romancero del siglo xvi.

95 Y un octavo repertorio abreviado sobre un facticio de impresos en catalán, que reúne la mayoría de los pliegos poéticos más antiguos en esta lengua y de la imprenta valenciana.

> de romancero en *open access*, dos de ellos de fuentes impresas y otro de fuentes manuscritas, en un arco cronológico que irá desde el siglo xv hasta 1552: el *Repertorio abreviado del romancero manuscrito* (RARM), el *Repertorio abreviado del romancero incunable* (RARI) y el *Repertorio abreviado del romancero impreso (1501-1552)* (RAR16). Este último, a cargo de Mario Garvin (2022), es el primero que se ha publicado en abierto y, en breve, lo harán los dos siguientes. Al *Repertorio abreviado de incunables poéticos* (RAIP) se unirá en el futuro un *Repertorio abreviado de postincunables poéticos* (RAPP), en ambos casos atendiendo a la tradición castellana como criterio de indexación. El matiz es oportuno porque también se contempla la tradición catalana, cuya prioridad será elaborar el *Repertori abreujat d'impresos del Natzaré* (RAINA), el volumen facticio de impresos poéticos breves más importante, por su antigüedad y rareza, de la tradición catalana, de la cual recoge su(s) primer(os) pliego(s) poético(s). Comparable, por tanto, a los *Pliegos Poéticos de Praga* castellanos por su singularidad y por su antigüedad, aquí aun mayor, que compensa el número menor de impresos. Este será el primer paso, antes de generar el *Repertori abreujat d'incunables amb poesia catalana* (RAIPC) y el *Repertori abreujat de postincunables amb poesia catalana* (RAPPC). (Martos, 2024a, p. 334)

Junto a estas aplicaciones en línea, disponemos de un rico repositorio de *Descripciones codicológicas* de fuentes manuscritas en continua expansión, con alrededor de cincuenta códices poéticos estudiados, así como una sección íntegramente dedicada al extenso corpus que suponen los 11 volúmenes del cancionero decimonónico MN13, el proyecto ilustrado de recopilación del *Cancionero general del siglo xv* (Martos, 2012a). Este se encuentra almacenado en una base de datos relacional a la que se le ha dotado de una interfaz que permite la búsqueda y listado de sus contenidos e índices (Díez Garretas *et al.*, 2012; Martos, 2018a). Asimismo, esta sección se completa con la descripción codicológica de dicho cancionero (Moreno, 2012a) y un estudio de sus fuentes manuscritas e impresas (Moreno, 2012b). Por otra parte, bajo el paraguas del mismo grupo de investigación, se creó una revista electrónica en el 2012, la *Revista de Cancioneros Impresos y Manuscritos* (RCIM), indexada en *Scimago*, *Scopus*, *Clarivate* y con el sello de calidad de la FECYT, que tiene una periodicidad anual y de la que, a día de hoy, se han publicado trece números.

Además de las correspondientes secciones en CIM, sobre humanidades digitales aplicadas al Romancero, contamos con un proyecto de gran alcance centrado exclusivamente en este género poético: el *Pan-Hispanic*

Ballad Project (PHBP), coordinado por Suzanne H. Petersen de la University of Washington. Este portal proporciona un conjunto de bases de datos interrelacionadas que dan acceso a la bibliografía de referencia, así como a un extenso corpus de textos que comienza en el siglo xv, con diversas entradas acompañadas de la interpretación oral o de la notación musical correspondiente. Petersen, antes de embarcarse en la coordinación de este repositorio, colaboró en el diseño de las primeras bases de datos de proyectos romancísticos con el Instituto Universitario *Seminario Menéndez Pidal* (SMP) en una línea de investigación que, actualmente, tiene su sede en la Fundación Menéndez Pidal[96]. Esta fundación tiene en marcha varios proyectos humanísticos digitales, entre los que cabe destacar, por su relación con el tema tratado, el de creación del *Archivo Digital del Romancero* (ADR) que, una vez finalizado, contendrá la digitalización del *Archivo del Romancero Menéndez Pidal-Goyri*, además de las colecciones recopiladas por Juan Menéndez Pidal, por José Amador de los Ríos y las copias y transcripciones de los originales de Marià Aguiló para el *Cançoner popular de Catalunya*.

Limitado a la tradición portuguesa, en esta misma línea encontramos la web *O Arquivo do Romanceiro Tradicional em Língua Portuguesa* (*Romanceiro.pt*), coordinada por Pere Ferré, de la Universidade do Algarve, que alberga una base de datos de los romanceros de la tradición oral moderna portuguesa publicados entre los siglos xix y xxi. Se ha convertido en una plataforma que acoge las investigaciones sobre romanceros en lengua portuguesa, como pone de manifiesto la incorporación a este portal de *Garret Online*, un proyecto que tiene como objetivo editar digitalmente el *Romanceiro* de Almeida Garret, el poeta romántico portugués, así como la colaboración con el *Archivo Digital del Romancero* de la Fundación Menéndez Pidal. De más reciente aparición es la web *Revisões literárias: a aplicação criativa de romances antigos (sécs. xv-xviii)* (*RELIT-Rom*), dirigida por Teresa Araújo, de la Universidade Nova de Lisboa, y que complementa a *Romanceiro.pt* con objetivos similares para los romanceros situados entre los siglos xv y xviii.

Las primeras aproximaciones desde las humanidades digitales a los estudios gallego-portugueses sobre poesía de *cancioneiro* van unidos, ineludiblemente, al *Centro Ramón Piñeiro para a Investigación en Humanidades*

96 El grupo de romanceros de la UCM sobre romancero panhispánico se ha disuelto por la jubilación de varios de sus miembros y los proyectos se han trasladado a la Fundación Menéndez Pidal según indica su página web en https://www.ucm.es/smenendezpidal/romancero-panhispanico [consulta: 29/04/2023].

(CRPIH), con sede en Santiago de Compostela, un referente indiscutible en este campo. Son responsables de la creación y mantenimiento de múltiples bases de datos, entre las que se encuentran las centradas en estudios medievales *Palmed*, coordinada por Mercedes Brea y Pilar Lorenzo y que ofrece la transcripción paleográfica de los testimonios manuscritos de la lírica gallego-portuguesa, y *MedDB*, también coordinada por Brea y Lorenzo, que almacena el corpus completo de las cantigas medievales de los trovadores gallego-portugueses. Pero los proyectos de este centro de investigación no solo se centran en este periodo, sino que abarcan hasta la actualidad, como lo demuestran *CORGA*, un corpus documental etiquetado automáticamente representativo del gallego actual, y *TERGAL*, un banco de datos terminológicos que recoge las denominaciones recomendadas para conceptos de las lenguas de especialidad.

Centrado exclusivamente en la poesía medieval gallego-portuguesa y coordinado por Manuel Ferreiro desde la Universidade da Coruña se encuentra *Universo Cantigas* (UC), con el objetivo de llevar a cabo la edición crítica digital de todos los textos de la lírica profana gallego-portuguesa. Es un proyecto que deriva del *Glosario da poesía medieval profana galego-portuguesa* (GLOSSA), que disponía de web propia y ahora se ha integrado en una sección de UC (Ferreiro, 2019, p. 1633). En este portal se pone en abierto un extenso corpus cancioneril representativo del gallego-portugués medieval, que se complementa con el proyecto de estudio de las *Cantigas de Santa Maria* (CSM) de la Universidad de Oxford:

> Composed and compiled in the late 13th century as a personal project of Alfonso X of Castile, it comprises 419 poems in medieval Galician-Portuguese, celebrating the Blessed Virgin Mary, to whom the King professed particular devotion. Most of the poems are set to music and many are richly illustrated. The entire collection is contained in four precious manuscripts (To, T, F, E) all produced in Castile before 1283, representing different stages in the conceptualization and execution of the King's project (Parkinson, 2019, p. 77).

En este caso, se dispone de una base de datos dotada de una interfaz de búsqueda que contiene los textos, los manuscritos y las miniaturas del repertorio de piezas líricas en gallego medieval compuestas en la corte de Alfonso X. Asimismo, dedica una sección al repositorio, que almacena las ediciones críticas o en borrador de los poemas en formato PDF, con enlaces a las correspondientes fichas de la base de datos.

En lo que respecta a la poesía medieval catalana, el *Repertorio informatizzato dell'antica letteratura catalana* (RIALC) fue un proyecto pionero en su campo, dado que su primera versión en línea apareció el 11 de agosto de 1999, con sede en la Università Federico II di Napoli, con el cual han colaborado a este efecto la Universitat Autònoma de Barcelona y la Universitat de Girona. Ha sido coordinado desde entonces por Costanzo Di Girolamo y co-dirigido junto a Lola Badia. RIALC está formado por el corpus de la poesía catalana de los siglos XIV y XV[97]. Ofrece la posibilidad de acceder a los textos mediante un índice de autores, de títulos, así como por *incipit*, si bien no permite búsquedas textuales. Aunque se completó en el 2001 y a pesar del reciente fallecimiento de Di Girolamo, sigue en línea y plenamente operativo. Por otra parte, su proyecto hermano, el *Repertorio informatizzato dell'antica letteratura trobadorica e occitana* (RIALTO), también coordinado por Costanzo Di Girolamo, recoge el corpus de la poesía trovadoresca y occitana, así como un corpus del occitano antiguo (CAO).

En el portal *NARPAN*[98], dedicado a la cultura y literatura catalana medieval, entre otras secciones, se da cabida a *Cançoners DB*. Esta web con entidad propia, dirigida por Miriam Cabré y Sadurní Martí, es la que,

97 En los siguientes términos: "Finalità del Rialc è l'inventario critico della poesia catalana dei secoli XVI e XV. Come è noto agli specialisti, ma assai meno al lettore medio, sia pure di buona o alta cultura, si tratta di una tradizione di notevole qualità, che prende vita autonoma sul finire della civiltà trobadorica (secoli XII e XIII), innestandosi in essa e continuandone dapprima la lingua, via via sempre più localizzata geograficamente, fino a giungere alle soglie della Modernità con il canzoniere di Ausiàs March, il capolavoro della lirica europea quattrocentesca, e con l'opera, permeata di forti tratti umanistici, di Joan Roís de Corella. Una più approfondita conoscenza della cultura letteraria catalana del tardo Medioevo, aperta all'Italia e alla Francia, oltre che memore delle sue radici occitane, rappresenta anche un indispensabile viatico per l'adeguata comprensione dei primi secoli di un'altra grande letteratura peninsulare, quella castigliana" (*RIALC*, s.f.).

98 *NARPAN* es un proyecto y un espacio digital conjunto de la Universitat Autònoma de Barcelona, la Universitat de Barcelona y la Universitat de Girona, formado por un grupo internacional de investigadores liderado por Lola Badia. En este caso, sus investigaciones se centran en la cultura y la literatura de la baja Edad Media, con aportaciones destacables al campo de las humanidades digitales catalanas como el *Corpus Digital de Textos Catalans Edat Mitjana i Renaixement* (CDTC), que se materializa en seis bases de datos consultables a través de la web del grupo: *Llull DB*, *Sciencia.cat DB*, *Translat DB*, *Arnau DB*, *Eiximenis DB* y *Cançoners DB*.

por su temática, entronca de manera más directa con este trabajo[99]. En ella, encontramos una base de datos relacional que permite la consulta de los autores, textos y bibliografía relacionada, así como una biblioteca digital, que hereda el trabajo realizado en RIALC para renovarlo, corregirlo y ampliarlo. No obstante, para acceder a la funcionalidad completa de *Cançoners DB*, el usuario debe de pasar por un proceso de registro previo que solicita su nombre y su dirección de correo, un requisito que no es necesario en RIALC. Otra de las secciones de interés en NARPAN es la dedicada a Cerverí de Girona, dirigida por Miriam Cabré, en la que se aportan datos sobre las obras y fuentes de este trobador, así como, en un futuro, la edición digital de su obra completa. En paralelo a este portal y con la misma temática trovadoresca, Cabré dirige la revista digital *Mot so razo* y el proyecto *TrobEU*. Esta revista comenzó su andadura en 1999, tiene una periodicidad anual y combina los artículos de investigación con la alta difusión en acceso abierto. Por su parte, el objetivo de *TrobEU* es profundizar en los canales de recepción histórica del legado de los trovadores en las cortes catalanas. Para ello, "un equip d'investigadors pertanyents a deu centres de Catalunya, Itàlia, França i Anglaterra han sumat esforços per avançar els nostres coneixements sobre dos aspectes fonamentals d'aquesta recepció: les corts que van acollir els trobadors i els manuscrits que van transmetre'n les obres" (*TrobEU*, s.f.).

Este interés científico por la poesía de cancionero, así como su atención desde las humanidades digitales, no se limitan en la Universitat de Girona a la Edad Media y, así, es muy destacable el espacio de referencia que supone el grupo *NISE*, dirigido por Albert Rossich y dedicado a la literatura catalana de la Edad Moderna, con especial atención a la poesía, como demuestran sus dos magnos proyectos sobre la obra completa de Francesc

99 Los objetivos de *Cançoners DB* quedan expuestos de manera abreviada en su página de entrada: "*Cançoners DB* és un portal temàtic i una base de dades sobre els cançoners catalans medievals. Ofereix descripcions i dades sobre els manuscrits, impresos, autors i obres, una bibliografia exhaustiva i una biblioteca electrònica de textos". Y, con mayor especificación, en su introducción: "*Cançoners DB* és una base de dades, biblioteca digital i portal dedicats, des de 2006, als cançoners medievals que transmeten poesia lírica i narrativa en català (c. 1250–1500). Es tracta d'un corpus d'unes 2500 peces escrites per 250 autors copiats en 30 cançoners, amb una tradició extravagant que inclou 40 altres manuscrits i edicions antigues. *Cançoners DB* combina, en un entorn PHP-MySql, quatre taules complexes amb informació sobre 1) Manuscrits i edicions, 2) Autors, noms de persona i toponímia, 3) obres i 4) bibliografia secundària" (*Cançoners DB*, s.f.).

Fontanella y de Vicent Garcia, aunque son muchos otros los poetas de los que se ofrece su edición digital. El grupo ha generado la web *NISE*, que da acceso a una base de datos que recoge la poesía catalana del Barroco en cancioneros y otros testimonios poéticos manuscritos e impresos, estudios sobre la literatura catalana moderna y ediciones críticas digitales. En este último caso, además, con una innovadora presentación visual que aprovecha las posibilidades del hipertexto para mostrar, de una forma no intrusiva y ordenada, las variantes, los comentarios a estas y las anotaciones, junto a cada línea del texto.

No cabe duda de que la tecnología informática permite crear nuevas formas de presentación de las ediciones críticas y las transcripciones que facilitan la labor investigadora. La Biblioteca Virtual Miguel de Cervantes (BVMC) ha sido, desde su nacimiento, incubadora de algunas de ellas. En lo que al medievalismo afecta, ha creado las bibliotecas de tres de los grandes escritores de la literatura catalana: Ramon Llull, Ausiàs March y Joan Roís de Corella. Estos espacios se enmarcan en la Biblioteca Lluís Vives y actúan a modo de repositorio de códices y/o incunables o post-incunables, así como de transcripciones, ediciones críticas e investigaciones sobre su vida u obra; sin embargo, destaca, por su novedoso concepto e intachable ejecución, la edición sinóptica de las poesías de Ausiàs March, dirigida por Llúcia Martín Pascual y Rafael Alemany Ferrer[100]. Es una herramienta que permite obtener, en una única página, la poesía objeto de estudio y los testimonios que se tienen de ella y que resultan de interés, lo que permite establecer rápidamente comparativas entre ellos que confirmen hipótesis (Martín Pascual, 2020, pp. 327-328, 2024, p. 114).

Las fuentes de la poesía de cancionero tienen un interés especial, quizás mayor que el de otros géneros, por el carácter antológico que a menudo presentan y la complejidad de formación de estos testimonios, no siempre en una misma fase o de una manera lineal. Es por ello que los proyectos digitales que contemplan esta perspectiva literaria son de especial interés, especialmente los centrados en las filologías hispánicas, como es el caso destacado de *PhiloBiblon*, con sus bibliografías BETA, BITECA y BITAGAP, dedicada a los textos literarios castellanos, catalanes, gallegos y portugueses, así como BIPA, sobre poesía española áurea. Fue pionero en las humanidades digitales dedicadas al hispanismo literario, partiendo de criterios organizativos y catalogadores, como los textos (*texid*) y sus fuentes (*manid*),

100 https://www.cervantesvirtual.com/portales/ausias_march/edicio_sinoptica/ [consulta: 23/04/2023].

así como muchos otros secundarios que facilitan su búsqueda cruzada. Las fuentes son manuscritas e impresas y no solo poéticas, pero sí que están catalogadas y tratadas con amplitud de datos bibliográficos y filológicos. Pese a que este metacatálogo está formado por tres bibliografías, tal separación no impide efectuar la búsqueda en todas ellas a la vez, compartiendo la misma interfaz de visualización de fichas y resultados. Las fichas son de las más completas y aprovechan las posibilidades del formato digital para crear, mediante enlaces, una compleja red que permite navegar entre los registros que guardan algún tipo de relación, así como con enlaces externos, localizaciones de ejemplares o ediciones facsimilares digitales.

El *Catálogo de obras medievales impresas en castellano* (COMEDIC), dirigido por María Jesús Lacarra en el marco del grupo *Clarisel* de la Universidad de Zaragoza, ofrece un modelo similar a *PhiloBiblon*, al menos en su punto de partida, aunque generando un extenso catálogo de las obras medievales impresas que abarca desde las últimas décadas del siglo xv hasta finales del siglo xvi, con una profundidad mayor de datos y una investigación específica para cada uno de los textos que no solo supera la mera catalogación, sino que implica un importante avance científico. Esta base de datos facilita el filtrado de su contenido mediante un potente buscador que acota los resultados por medio de múltiples campos. A pesar de que no es un catálogo específico de obras poéticas, muchas de sus entradas sí que lo son y, asimismo, documentan sus fuentes concretas.

El *Servidor Web de Literatura Española Parnaseo*, al que nos hemos referido anteriormente, presta atención especial a la imprenta. Destaca en cuanto a los intereses de esta monografía el *Portal Celestinesco*, dirigido por Marta Haro y, como su nombre indica, centrado en *La Celestina* —una obra con poesía incorporada, no lo olvidemos (Deyermond, 1997)—. Este portal dispone de un catálogo de incunables e impresos del siglo xvi, con la virtud, además, de ofrecer imágenes de los testimonios originales que sirven de perfecto acompañamiento a las descripciones textuales, así como una sección en abierto dedicada, de manera específica, a los grabados, lo que supone una importante novedad en el campo de las humanidades digitales.

A estos proyectos específicos de literatura medieval e imprenta habría que sumar los grandes catálogos internacionales de incunabilística en línea. El *Gesamtkatalog der Wiegendrucke* (GW), con sede en la *Staatsbibliothek zu Berlin,* es la versión en línea del catálogo en papel que se viene editando desde 1925 y que recopila los impresos del siglo xv alfabéticamente. Su versión en línea, con unas 36 000 descripciones de incunables, aprovecha las posibilidades del formato digital añadiendo información complementaria

respecto a la versión física. Dado que la edición en volúmenes se está publicando por estricto orden alfabético, la versión web les permite incorporar las ediciones que se conocen con posterioridad a la impresión del correspondiente volumen. En cada una de sus fichas se incluye el autor, título, lugar de impresión, impresor y fecha, así como una colección de entradas descriptivas de la obra y la localización de los ejemplares. Gran parte de estos campos enlazan a otras fichas lo que permite, de una forma sencilla, la navegación por los distintos registros de la base de datos a partir del mismo autor o impresor. Es solidario con el propio ISTC, del que se hablará a continuación, y, en un espíritu científico de colaboración, enlaza a sus fichas directamente, de la misma manera que lo hace a las digitalizaciones de ejemplares concretos disponibles en línea. Viene completado por un proyecto paralelo, con sede en la misma institución, como es el *Typenrepertorium der Wiegendrucke* (TW), un catálogo de tipografías de incunables, con el que va enlazando automáticamente el GW en cada una de sus fichas. Una de las grandes novedades de este catálogo es la incorporación de las tipografías de talleres concretos en formato electrónico, a partir de los tipos móviles utilizados en los impresos, con la posibilidad de descargarlas e instalarlas localmente, así como las imágenes escaneadas de las fichas manuscritas creadas por los investigadores antes de la existencia de la web, dado que contienen información adicional no estructurada con un difícil encaje en los campos de la ficha digital. Esta herramienta facilita el análisis tipográfico de los incunables poéticos y es, de hecho, uno de los principales recursos en línea que contribuyen al estudio de su materialidad.

Con un propósito similar al del GW, con cuyas fichas también enlaza, pero desarrollado por la *British Library*, tenemos el *Incunabula Short Title Catalogue* (ISTC), el cual mantiene un registro en línea de todos los incunables impresos con tipos móviles. Cada una de sus entradas incluye los autores, el título, la lengua del texto, el impresor, el lugar y fecha de impresión, el formato y la localización de las copias. Asimismo, se proporcionan accesos directos a sus correspondientes facsímiles digitales en caso de existir.

El *Universal Short Title Catalogue* (USTC), con sede en la University of St Andrews, abarca desde la invención de la imprenta hasta finales del siglo XVI, con un total de 4 millones de copias registradas, lo que lo convierte en uno de los mayores catálogos en línea al respecto. Permite acotar el resultado de las búsquedas por múltiples campos, acompañados de un gráfico que muestra visualmente el número de copias por años. Aunque los registros no están enlazados entre ellos internamente, sí que referencian a las fichas de otros catálogos externos, como el GW y el ISTC. En contraste, el *Iberian*

Books (IB) es un proyecto de la *School of History* del *University College Dublin* que se centra en los libros publicados en la península ibérica y el Nuevo Mundo entre 1472 y 1700, dispone de unas fichas hipertextuales que permiten navegar entre sus registros de una forma intuitiva, acompañadas de un completo motor de búsqueda que facilita acotar con precisión los resultados. Además, es de destacar la incorporación de la geolocalización de cada uno de los ejemplares sobre un mapa interactivo.

La *Bayerische StaatsBibliothek* en Múnich posee la mayor colección de incunables del mundo, con más de 20 337 ejemplares, por lo que su catálogo en línea en forma de base de datos, el *Bayerische StaatsBibliothek Inkunabelkatalog* (*BSB-Ink*), que permite su consulta a través de un total de doce campos de filtrado distinto, se convierte en un referente importante para el estudio internacional de la imprenta en el siglo XV. Sin embargo, el resultado de la búsqueda es poco amigable, dado que únicamente presenta un listado de sus números de registro que enlazan a la correspondiente ficha, que muestra el título, impresor, lugar de impresión y fecha, la descripción material y su proveniencia, así como los correspondientes enlaces al GW y al ISTC y, si existe, su facsímil digital.

Otra de las bibliotecas con una colección de incunables es la Bodleiana de la Universidad de Oxford, que también los tiene registrados en su conocido catálogo en línea: el *Bod-Inc online*. La consulta de sus entradas puede filtrarse mediante múltiples campos y el resultado muestra un listado ordenado por su número de registro con el autor, título y datos de la impresión. Adicionalmente, permite acotarlo por autor, impresor o lugar de impresión. Cada una de las fichas dispone de los datos básicos a los que añade un breve análisis de contenido, su estructura colacional, referencia a otros catálogos, y una descripción detallada de la copia que reside en sus fondos.

En España, el Ministerio de Cultura y Deporte ha puesto en marcha el *Catálogo Colectivo del Patrimonio Bibliográfico Español* (CCPB), que permite buscar entre todos los fondos que poseen las bibliotecas e instituciones españolas, tanto públicas como privadas, desde la Edad Media hasta la actualidad. Tiene la posibilidad de consultar el catálogo completo o acotarlo por comunidades, ciudades o, incluso, localizaciones concretas. Cada una de las fichas está formada por una descripción de la edición, así como la localización de distintos ejemplares. Destaca la posibilidad de exportar los registros a formatos de intercambio estándar, como ISDB, MARC, MARCXML o DublinCore, lo que facilita la inclusión de su contenido en aplicaciones externas. Esta política aperturista de datos también se está realizando en otros catálogos, como en el caso ya aducido de *PhiloBiblon*, que

está inmerso en una profunda transformación para adaptarse a un formato de edición colaborativo.

> When will all this happen? We are currently prototyping *PhiloBiblon* on *FactGrid* during the 2021-2022 academic year, with a follow-on implementation grant to start –if the application is successful– in 2023. Since a logical and conceptual data model already exists in the *PhiloBiblon* schema, we have focused during this prototyping period on mapping it to the *Wikibase* model. (Faulhaber, 2022, p. 193)

Finalmente, el *Material Evidence in Incunabula* (MEI) es una base de datos alojada y mantenida por el *Consortium of European Research Libraries* y que registra los rasgos materiales que se conocen sobre los libros impresos del siglo xv, como poseedores, encuadernación, anotaciones manuscritas, sellos o precios. En la creación y mantenimiento de sus registros, alrededor de 50 000 actualmente, colaboran tanto librerías europeas como americanas. Sigue un novedoso enfoque basado en la aplicación de la localización geográfica a lo largo de los años que permite rastrear el movimiento de cada una de las copias.

En definitiva, todos estos proyectos son una demostración, por tanto, de que las humanidades se han beneficiado de la informática, especialmente en lo que respecta a la investigación en poesía medieval y, de manera complementaria, en cuanto a los *instrumenta* que permiten el estudio de sus fuentes impresas. Filología e informática se han unido en una perfecta simbiosis para evolucionar y beneficiarse mutuamente. Las computadoras ahora son capaces de entender el lenguaje humano gracias a la filología y los textos, verdaderos transmisores de cultura y conocimiento, se han deslocalizado y son accesibles a golpe de clic.

3. Digitalización y difusión de textos

La migración de lo físico a lo digital ha llevado aparejada, inevitablemente, un cambio en la metodología de trabajo. Desde finales del siglo xx, con la extensión del uso de Internet, el papel ha ido perdiendo paulatinamente el privilegio que había mantenido desde la invención de la imprenta en cuanto a principal difusor del conocimiento[101]. La ventaja más evidente de la digitalización es el fin de la dependencia del contenido con el soporte —texto y papel—, ya que con este proceso, la obra pasa de estar confinada en un lugar a ser accesible desde cualquier punto con conexión a Internet, siempre que se difunda a través de este medio.

101 Durante "unos 550 años ha dominado la tecnología de la imprenta, y continúa con nosotros, adaptándose con esfuerzo a los nuevos tiempos; pero en los últimos 25 años, la relación entre intelectuales y ordenadores primero y luego la generalidad del uso de Internet, ha cambiado nuestros cerebros y revolucionado las prácticas intelectuales y ha modificado nuestras costumbres y hábitos sociales" (López Poza, 2014, p. 151).

3.1. La necesidad de digitalizar

La digitalización de los impresos y manuscritos se ha convertido, en muchos casos, en una necesidad. La oxidación, acidificación o desmagnetización, entre otros procesos degenerativos, hacen desaparecer las obras en formato físico y, aunque las técnicas de restauración pueden ralentizar su deterioro, en ningún caso lo logran detener (Tacón Clavaín, 2008, p. 29). Si nos centramos en los libros, su integridad puede verse afectada por diversos factores, tanto intrínsecos, provocados por los materiales con los que están confeccionados[102], como extrínsecos, causados por agentes ajenos al proceso de fabricación, normalmente asociados a un incorrecto almacenamiento[103].

Durante la Edad Media y hasta la segunda mitad del siglo xv, "el pergamino es el soporte de escritura predominante" (Pedraza Gracia, 1997, p. 14)[104], cuya confección en piel orgánica ofrece una superficie no homogénea, con zonas de distinto grosor, que hacen que absorba la humedad del ambiente de forma irregular. Esta característica inherente, con el paso del tiempo, termina provocando deformaciones, grietas, arrugas y roturas (Vergara, 2002, p. 75). No obstante, pese a ello, es un material más duradero que el papel que, sin embargo, con el nacimiento de la imprenta se impuso por su menor coste y facilidad de fabricación (Lacarra, 2019, p. 295), con lo que el uso del pergamino quedó relegado a tiradas lujosas (Martos, 2019, p. 694).

Si bien se extiende su uso en Europa a partir de las últimas décadas del Medievo, la invención del papel debe datarse hacia finales del siglo i y se suele atribuir a T'sai Lun, un eunuco que vivió en la corte imperial durante el reinado del emperador Hedi. Hasta ese momento, el soporte tradicional de escritura en China era la seda, que era muy costosa, por lo que confeccionó un nuevo material a partir de materias primas más baratas, como

---

102 Para una enumeración exhaustiva de los factores intrínsecos de degradación de los libros véase Forniés Matías & García Quiroga, 2014.

103 Sánchez Herrador *et al.* (2010) muestran, con imágenes capturadas del fondo de la Biblioteca Pública del Estado-Provincial de Córdoba, el resultado del deterioro producido por factores extrínsecos en los libros.

104 "Aunque históricamente el papiro fue empleado en las cancillerías europeas y en la corte papal para redactar algunos documentos hasta al menos el siglo xi, su uso como soporte para la copia de obras no va más allá del siglo vi o vii, y ya de forma residual" (Avenoza, 2019, p. 70).

cortezas vegetales, fibras de morera y restos de tejidos[105]. No obstante, tuvieron que pasar mil años para que llegase, finalmente, a Europa a través de los árabes (Dahl, 1982, pp. 41-43), que lo venían utilizando desde el siglo VIII, se supone que a partir de la captura de unos fabricantes chinos durante la batalla de Talas en el año 751 (Bloom, 2001, p. 42). Su introducción en la península ibérica se estima que debió de ser a mediados del siglo X, a través de Córdoba, si bien no está atestiguado. Existe una leyenda que sitúa un molino papelero en Xàtiva en el año 1056[106], aunque las primeras referencias documentales que se tienen de este emplazamiento son de mediados del siglo XII (Balmaceda Abrate, 2008, p. 104; Valls i Subirà, 1978, I, p. 133)[107]. En el mismo siglo, se sitúa en Cataluña un molino de papel junto al río Besós, concretamente, en el año 1113, uno en Sant Vicenç de Jonqueres en el 1158 y uno en La Riba en el 1159 (Valls i Subirà, 1978, I, p. 152). Un siglo después, en 1276, aparecen los molinos de la zona de Fabriano y, años más tarde, en Bolonia, Padua y Génova[108].

Para su confección, los primeros papeles asiáticos se fabricaban poniendo los vegetales escogidos en un baño de cal con el fin de ablandarlos. Una vez habían fermentado, se trituraban, maceraban y la pulpa resultante se

---

105 Aunque es indudable que la invención del papel se produjo en China, no están claras las fechas ni la forma en la que se produjo, lo que convierte la historia de T'sai Lun en una leyenda, dado que se han encontrado diversas muestras de papel que hacen pensar que su invención se produjo dos o tres siglos antes (Bloom, 2001, p. 32). De hecho, "la tradición china lo atribuye a un hombre llamado Han Hsin, que vivió entre los años 247 y 194 a.C." (Vergara, 2002, p. 15).

106 "Una de las primeras citas relacionadas con el papel menciona a un tal Abu-Masafya, quien elaboraba este producto en el año 1056 'junto a la vieja acequia' con más de veinte operarios. Su hijo mayor huye, cuando el Cid conquista Valencia, y funda otra fábrica en Ruzafa. En 1085 se alza otra en Toledo. Tenemos también el testimonio de Pierre le Vénérable, abad de Cluny (1091-1156), el cual describe en tono despectivo unos ejemplares del Talmud, hechos a base de trapos: *ex rasuris veterum pannorum*, vistos durante su peregrinación a Santiago de Compostela" (Ruiz García, 2002, pp. 64-66).

107 Los historiadores del papel no están de acuerdo en la fecha concreta. "Játiva documentada y alabada la calidad de su producción papelera por el geógrafo *al-Idrīsī* en 1154 continúa siendo el centro papelero más conocido y prestigioso de la época. *al-IDRĪSĪ, Geografía de España*, Anubar ediciones, Valencia, 1974" (Balmaceda Abrate, 2008, p. 104). "La fecha que nos da *al-Idrīsī* de la elaboración del papel en Xátiva, se puede colocar aproximadamente en el año 1147" (Valls i Subirà, 1978, I, p. 133).

108 Así como en la zona de España no evolucionaron tecnológicamente, los italianos introdujeron varias mejoras que aumentaban la calidad del papel. Entre ellas, destaca el uso de mazos de hierro tachonados que trituraban más eficazmente los vegetales (Bouyer, 1994, p. 6).

mezclaba con agua. Cada una de las hojas se creaba sumergiendo en esta pasta un tamiz de fibras de bambú o de tela que tenía la forma y tamaño deseado y que permitía extraer una fina capa, para volcarla sobre paños y prensarla, a fin de eliminar el exceso de agua, tras lo cual se ponía a secar. Para darle lustre, se frotaba con una piedra lisa y, finalmente, se le aplicaba una cola a base de algas o de savias que le confería cierta impermeabilidad y resistencia. Este método primitivo fue evolucionando con el tiempo: los árabes incorporaron el reciclado del trapo –al que debe el nombre con el que se le conoce: papel de trapo–, los tamices con malla metálica y la pasta de almidón de harina de trigo para el encolado (Asunción, 2009, pp. 29-33).

> Los manuscritos hispánicos más antiguos con cuadernos que incluían folios de papel árabe (junto a otros de pergamino) son, precisamente, dos códices del siglo xi procedentes del Monasterio de Santo Domingo de Silos, un *Glossarium latinum* (París, BnF Nouvelles Acquisitions Latines 1296)[109] y un *Brevarium gothicum seu mozarabicum* hoy en Silos, Biblioteca de la Abadía de Santo Domingo de Silos, BASDS códice 6. (Avenoza, 2019, p. 80)

Obviamente, cada maestro papelero aplicaba su fórmula y técnica particular que determinaba el grosor, textura y calidad final del papel. Por ello, era una cuestión de tiempo que cada molino quisiera introducir una señal en sus productos que lo identificase como el proveedor, pero que, al mismo tiempo, no fuese perceptible a simple vista para no interferir en la lectura ni restar espacio de escritura: esta función la cumplió la marca de agua o filigrana. La primera que se conoce, la cruz griega, se crea en la citada zona de Fabriano y aparece en un documento fechado en Bolonia sobre el 1282 (Briquet, 1907, p. 325)[110]. Este elemento identificador del papel se generaba con alambres que se colocaban sobre el tamiz, lo que provocaba que en esa zona se concentrase menos pulpa y, con ello, quedase

---

109 Se puede consultar una digitalización a partir del original en https://gallica.bnf.fr/ark:/12148/btv1b84559374/f1.item [consulta: 17/05/2023]. Tal como apunta Avenoza (2019, p. 80), en su ficha bibliográfica aparece catalogado como perteneciente al siglo xii y con el título *Autre glossaire, en écriture visigothique, du XIIe siècle*.

110 Stevenson (1968, p. 68) apunta que existen dudas sobre el año exacto del documento, dado que la filigrana es idéntica a la de otro documento fechado en 1294 catalogado por Piccard (1961-1997). La versión digital se puede consultar en https://www.piccard-online.de/detailansicht.php?klassi=011.002.001.001&ordnr=162496&sprache=en [consulta: 25/05/2023].

marcada la filigrana mediante cierta transparencia. Con esta modificación, se conseguía imprimir en el producto el sello del fabricante de forma invisible, a no ser que se pusiese al trasluz, sin necesidad de reservar un espacio físico para ello, con lo que, en realidad, se había inventado una suerte de publicidad encubierta que permitía identificar el proveedor del papel ante el reconocimiento de su calidad[111]. El uso de filigranas se extendió rápidamente entre los artesanos de la época y, actualmente, son un método que ayuda a establecer el origen y las fechas de los códices y de los impresos, con repercusiones en la propia redacción de las obras (Díaz de Miranda, 2012, p. 71), todo ello con las reservas obvias.

Estos métodos artesanales de fabricación del soporte de la escritura lo dotaban de una resistencia al paso del tiempo, sobre todo en el caso del papel de lino con el que están confeccionados los incunables[112], que, a partir del siglo XVIII, disminuyó, con la introducción de la pila holandesa y aditivos químicos: el hipoclorito para blanquear, el alumbre para endurecer la cola y la colofonia para dotarlo de resistencia (Crespo Nogueira, 1992). Con el paso del tiempo, se produjo una degradación de la calidad que también se vio reflejada en la tinta. Las tintas medievales utilizadas en los manuscritos se componían básicamente de tres ingredientes: una sal metálica, ácido tánico[113] y goma arábiga como aglutinante (Díaz Hidalgo

111 Existen tres teorías sobre el origen y significado de las filigranas: la primera sostiene que fueron creadas por los Albigenses como una forma secreta de comunicación, la segunda establece que llegaron a los moldes de forma accidental, y la tercera, actualmente la más aceptada, defiende que es la marca distintiva del papelero y que este adoptaba un signo para ser identificado (Hidalgo Brinquis, 1992, pp. 193-194).

112 "En las ocho pruebas estudiadas, con tres muestras cada una, representadas por veinticuatro microfotografías, se han escogido papeles de todas clases, desde los de mejor calidad, hasta los que necesitarían ser rechazados como papelote. Todas las fibras provienen de lienzos de lino, sin ninguna excepción, no hallándose, ni el más pequeño rastro de algodón, por el que tendremos que esperar los dos próximos siglos para que pase de ser una fibra completamente desconocida en los tiempos que estudiamos, a pasar a competir con el lino. Así que se han estudiado muestras tales como las de las hojas del *Lamentio yprocras* de Arnau de Vilanova, impreso en Valencia en 1495. El *Missale Caesaraugustanum*, de Zaragoza en 1498, hasta el *Tratado de Officio*, Toledo 1568, y la *Suma de Varones* de 1590, hasta la *Crónica del muy esclarecido Príncipe y Rey don Alonso el onzeno*…, también impreso en Toledo en 1595" (Valls i Subirà, 1978, II, p. 243).

113 "De los tipos de taninos que existen, son los denominados hidrolizables los que se utilizaban para fabricar tinta negra, concretamente el ácido gálico. Éste se obtiene a partir de una excrecencia vegetal denominada 'agallas' o 'nuez de agallas', que se forma en la corteza de algunos árboles, como la encina, el roble o el alcornoque" (Pellón *et al.*, 2004, p. 46).

et al., 2018, p. 3). Al ser de tipo metalogálico[114] conseguían un alto grado de fijación, por lo que se consideraban superiores a las basadas en carbón, propensas a desprenderse[115]; sin embargo, son químicamente inestables y según los porcentajes utilizados, los taninos pueden llegar a reaccionar con el sulfato ferroso y crear ácido sulfúrico que, al contacto con el papel, lo desintegra y produce agujeros en las grafías (Vergara, 2002, p. 30). Esta pérdida de material provoca lagunas textuales y afecta a la integridad del soporte, que se vuelve quebradizo y acaba rompiéndose (Martos, 2016, p. 185); sin embargo, este deterioro no lo sufren las tintas utilizadas en las primeras prensas medievales, basadas en las fórmulas al óleo utilizadas por los pintores flamencos, cuya mezcla de componentes les han permitido perdurar en el tiempo sin afectar al soporte (Bloy, 1967, pp. 2-3)[116].

El método de fabricación del soporte de los incunables, así como la tinta utilizada en su impresión permiten su pervivencia durante siglos, pero es indudable que, para ello, deben de estar en condiciones óptimas de conservación y preservación sin renunciar a su difusión.

> La conservación tiene por objeto garantizar la transmisión de un objeto en el mismo aspecto —forma, contenido— en que este ha llegado a nosotros a través de actuaciones que eviten la alteración de sus materiales y su función. Se trata, en definitiva, de medidas para evitar la disminución cuantitativa y cualitativa de los elementos de una obra. (Allo, 1997, p. 279)

---

114 "La tinta negra, la ferrogálica, era la preferida y se conocía en la Edad Media con el término latino *encaustum*, precisamente por su composición ácida y por su poder de corrosión de algunos soportes caligráficos, que no se limitaban al papel, puesto que también llegaba a dañar el pergamino" (Martos, 2016, p. 180).

115 "Elles ne sont sujettes ni à la réduction, ni à l'oxydation; elles ne contiennent aucune substance dangereuse pour le support. Tutefois, le mélange liant/noir de carbone ne pénètre pas toujours très bien dans les fibres du papyrus, du parchemin ou du papier, et un grattage suffit souvent à faire disparaître l'écriture" (Zerdoun Bat-Yehouda, 1983, p. 15).

116 Existen documentos con listas de compra de los impresores que se han asociado a los materiales utilizados para crear la tinta (Bloy, 1967, pp. 5-6). Aunque se supone que cada uno utilizaba su fórmula particular, los análisis mediante fluorescencia de rayos X no han permitido establecer un patrón que permita diferenciarlos (Mommsen *et al.*, 1996, p. 355).

Las técnicas aplicadas para ello han ido variando a lo largo del tiempo. Se han encontrado evidencias en las excavaciones arqueológicas que confirman que en el mundo antiguo ya se aplicaban ciertos mecanismos, en primer lugar, para asegurar la estabilidad del soporte mediante la elección adecuada de los componentes y su tratamiento y, en segundo lugar, para prevenir las ya conocidas plagas bibliófagas[117]. Desde la época romana, se establece la orientación de las construcciones hacia el Este como método de conservación, dado que disminuye la humedad. En la Edad Media, además de avanzar en el tratamiento de las plagas[118] y ante la fragilidad de los textos escritos con tintas al carbón, se comenzaron a utilizar las tintas ferrogálicas por su mayor poder de fijación, un cambio que pretendía aumentar el nivel de conservación de los documentos y que, paradójicamente, provocó importantes pérdidas documentales por el proceso de oxidación que desencadenaban.

> Las copias de documentos por razones de seguridad, renovación o reparación —copias *ex caducitate*— también deben ser consideradas medidas de conservación. Los ejemplos a lo largo del periodo son numerosísimos. Tres privilegios en papiro del monasterio de Nonántola, concedidos por los Papas Adriano I, Juan VIII y Marino fueron mandados copiar en pergamino por la Cancillería pontificia de Inocencio II porque "ea quae de ipsis scriptis papiriis, ex quadam parte prae nimia vetustate consuntis colligere potuit in publicam formam redigere procuravit"; o también el caso de Inocencio III, que renueva un privilegio en papiro del papa Agapito II del año 946 porque "quasi iam nimia vetustate consumta, cum fuerint, non in pergameno, sed in papyro conscripta". En otras ocasiones no será el deterioro de antiguos soportes sino la desconfianza ante los nuevos la causa que determine la realización de estas copias; un buen ejemplo lo constituye el decreto de Federico II de Alemania en 1221, ordenando copiar en pergamino todos aquellos documentos originales realizados en papel que tuvieran que ser conservados más de un año en su cancillería. (Allo, 1997, pp. 259-260)

---

117 Se introducía el documento en una caja de madera con características repelentes e insecticidas, impregnándola con sustancias con propiedades similares y acompañándolo con plantas aromáticas (Allo, 1997, p. 258).

118 Se empezaron a utilizar nuevos productos insecticidas que se esparcían sobre los documentos. Eran unas mezclas que contenían derris y pelitre (Kraemer Koeller, 1973, I, p. 580).

Ya en la Edad Moderna, con el auge de las bibliotecas y archivos, se establecieron instrucciones con las medidas a adoptar contra robos, incendios, humedades, polvo y bibliófagos. Entre las actividades para la conservación de los documentos ya figuraba explícitamente la realización de copias. De hecho, en algunos casos, estas se utilizaban de salvaguarda, dado que, una vez generadas, se empleaban para la consulta y se guardaba el original en un lugar distinto[119]. A partir del siglo xviii y, en mayor medida, desde el xix, se extiende este método de reproducción como mecanismo para nutrir el fondo de las bibliotecas privadas y públicas[120].

> La copia completa de un impreso se produce cuando, por lejanía en el tiempo o en el espacio, o bien por cualquier otro motivo, incluso económico, no es posible adquirir el ejemplar directamente. (Martos, 2011, p. 209)

Dado que las bibliotecas poseían ejemplares que eran únicos e irreemplazables, comenzaron a proliferar los procedimientos de restauración para reparar los daños con técnicas y productos que, en ocasiones, han afectado

---

119 "En los archivos, la copia de los documentos más importantes para evitar su deterioro o la realizada a documentos ya deteriorados, siguió siendo una de las actividades conservadoras más destacadas. Así aparece dispuesto en las *Instrucciones filipinas para el archivo de Simancas* (1588), en las que además se obliga a guardar originales y copias en lugares diferentes y a utilizar las copias y no los originales para acceder a la información" (Allo, 1997, p. 262).

120 "Les relacions entre erudits que intercanviaven els seus còdexs per fornir les biblioteques personals a través de còpies manuscrites dels originals havia arribat també a les col·leccions públiques, els directors de les quals se sol·licitaven mútuament reproduccions de materials a fi de completar els seus fons" (Martos, 2012b, p. 135). Así lo ejemplifica el caso de la copia manuscrita del cancionero *H* de Ausiàs March (Zaragoza, Biblioteca Universitaria, ms. 210) y la transcripción italiana de la edición *c* de este mismo poeta (Barcelona, Carles Amorós, 1545): "En aquest context de còpies de còdexs medievals i, en concret, de cançoners transcrits totalmente o parcialment, que comença a mitjan segle xviii i que té el seu apogeu al començament de la segona metitat del xix (Campa Gutiérrez, 2006), en aquest marc d'intercanvi de sol·licituds entre erudits i entre bibliotecaris amb la finalitat de completar les seus col·leccions particulars o públiques, es va generar la còpia vuitcentista del *Cançoner de Saragossa*, que no és l'única de textos d'Ausiàs March" (Martos, 2012b, p. 137); "Sin duda, fue Giraldi quien encargó la copia manuscrita escurialense para un uso personal, generando así un *codex descriptus* con poca utilidad ecdótica, pero de gran interés para reconstruir los contextos de recepción de la obra de Ausiàs March en Italia" (Martos, 2014, p. 277).

a su futura conservación[121]. Ante esta evidente degradación material que se producía en algunos casos, a finales del siglo XIX surgió el concepto de *conservación preventiva* que aún permanece hasta hoy en día y que, en primera instancia, estudia los agentes que producen el deterioro con el fin de actuar antes de que ocurra. Son un conjunto de técnicas y procedimientos que vivieron su impulso a partir del siglo XX con los avances científicos y la sistematización de las medidas a llevar a cabo y que se siguen aplicando en la actualidad; sin embargo, el libro es un objeto frágil que se degrada con facilidad, de hecho, cada lectura, total o parcial, significa una alteración del mismo. Se podría decir que del mismo modo que un sitio de Internet cuenta las visitas que recibe, el libro guarda, de alguna manera, un recuerdo de cada uso que se ha hecho de él, con la diferencia de que el libro tiene un número limitado de usos o visitas (Sánchez Herrador *et al.*, 2010, p. 282).

Aunque las bibliotecas y archivos apliquen de forma sistemática métodos de conservación para evitar los factores extrínsecos de deterioro del material impreso[122], como el control de la humedad, la temperatura, la luz, el polvo y, por supuesto, las plagas (Tacón Clavaín, 2008, pp. 30-51)[123], estas medidas preventivas no consiguen detener totalmente su degradación, únicamente la ralentizan. Es por ello que, con el fin de traspasar las limitaciones de la conservación, surge la preservación, un concepto que incluye todas las consideraciones gerenciales y financieras, así como pautas para almacenamiento y ubicación, niveles de personal, políticas, técnicas y métodos aplicables a la preservación de los materiales de archivos y bibliotecas y a la información que contienen (Dereau & Clements, 1986, p. 5).

Entre estas técnicas para salvaguardar el contenido de los soportes físicos se encuentra una de especial importancia: la digitalización. Con ella, se crea una copia virtual que está exenta de los problemas de materialidad que tiene su representación física y, como valor añadido, permite su reproducción y difusión a través de las redes de comunicaciones. Es sobradamente conocido el ahorro económico y temporal que supone en los campos de la educación y la investigación la posibilidad de obtener el

---

121 Allo (1997, p. 262) refiere el uso de adhesivos, productos para eliminar manchas y sistemas para blanquear el papel y reavivar tintas.

122 Para una descripción detallada de las medidas a adoptar para cada uno de los factores de deterioro y las condiciones ideales véase Tacón Clavaín, 2008.

123 Aparte de estos factores que se aplican a los libros, los soportes magnéticos, como los disquetes y las cintas, se ven afectados por los campos magnéticos.

original digitalmente y, ante su inevitable pérdida, es un mecanismo que permite que generaciones futuras puedan consultarlo.

Ante esto, no es de extrañar que desde las instituciones se impulsen iniciativas que busquen en la digitalización una forma de preservar y difundir el patrimonio cultural. Uno de los primeros programas a nivel europeo se encuentra en la Decisión 2001/48/CE del Consejo de la Unión Europea para poner en marcha la iniciativa *eContent*[124], que tenía, entre sus objetivos, incrementar la disponibilidad y uso de contenidos digitales en la red, fomentando su diversidad cultural y multilingüismo, al tiempo que se creaban las condiciones favorables para su comercialización, distribución y uso. Esto se traducía en una aportación de fondos para la creación de repositorios digitales y la mejora de las oportunidades comerciales de las empresas de generación de contenidos digitales. Cuatro años después, se ampliaba el programa con *eContentplus*, enmarcado en la Decisión 456/2005[125]. En su exposición de motivos, se destacaba la importancia de digitalizar los formatos físicos tradicionales de transmisión de información, como el papel y las cintas magnéticas, así como el uso de Internet como plataforma para su difusión. En sus sucesivas convocatorias de financiación, se premiaban los proyectos relacionados con las bibliotecas digitales y, de hecho, el proyecto más importante que se financió fue la *Biblioteca Digital Europea*, bautizada finalmente como *Europeana*. Este agregador era la estrella del marco estratégico europeo *i2010*, que buscaba facilitar la convergencia al mundo digital de los países miembros (Alvárez & Vives, 2009, pp. 31-43)[126]. Con este proyecto, se armonizaban las iniciativas de digitalización surgidas en diversos países y se unían en un único portal en la red, un lugar creado para ser un punto de difusión y promoción de la cultura europea.

Por supuesto, este impulso europeo a la digitalización tuvo su reflejo en España. En la disposición adicional tercera sobre fomento de la difusión de obras digitales de la Ley 23/2006, de 7 de julio, por la que se modifica el texto refundido de la Ley de Propiedad Intelectual, aprobado por el Real Decreto Legislativo 1/1996, de 12 de abril, se establece lo siguiente:

---

124 https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32001D0048 [consulta: 28/05/2023].

125 https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32005D0456 [consulta: 28/05/2023].

126 Se puede consultar la comunicación completa en https://eur-lex.europa.eu/ES/legal-content/summary/i2010-information-society-and-the-media-working-towards-growth-and-jobs.html [consulta: 28/05/2023].

> El Gobierno favorecerá la creación de espacios de utilidad pública y para todos, que contendrán obras que se hallen en dominio público en formato digital y aquellas otras que sean de titularidad pública susceptibles de ser incorporadas en dicho régimen, prestando particular atención a la diversidad cultural española. Estos espacios serán preferentemente de acceso gratuito y de libre acceso por sistemas telemáticos, mediante estándares de libre uso y universalmente disponibles[127].

Aunque con esta ley se promovía el apoyo institucional del poder ejecutivo en este campo, las colecciones digitales en nuestro territorio ya venían financiándose desde finales de los 80. Concretamente, en "1986 se había iniciado en España el proyecto del *Catálogo Colectivo del Patrimonio Bibliográfico* con muy buenos resultados" (Agenjo & Hernandez, 2020, p. 2). Asimismo, la Subdirección General de Coordinación Bibliotecaria contaba desde el 2005 con *Hispana*, un recolector de repositorios con posibilidad de intercambiar datos en el estándar OAI-PMH (Carrato Mena, 2014, p. 17), hecho que facilitó el cumplimiento de los requisitos europeos que solicitaban la creación de un agregador nacional y su integración con *Europeana* dentro del proyecto *EuropeanaLocal*[128], financiado por los fondos de *eContentPlus* (Martínez-Conde, 2012, p. 141)[129].

Pero más allá del sentido preservador y difusor de la digitalización que promueven estas iniciativas, se abre la posibilidad del tratamiento informático mediante la aplicación de algoritmos computacionales a las imágenes, una rama de la informática conocida como *visión por computador*. Los investigadores de este campo desarrollan técnicas matemáticas tan variadas como reconocer personas en una fotografía (Tolba *et al.*, 2006), detectar el estado de humor en el rostro (Revina & Emmanuel, 2021) o, algo que se lleva haciendo desde hace años: extraer la información de una matrícula (Poon *et al.*, 1995). Son aplicaciones que ya están en el mercado y se utilizan ampliamente, aunque aún se está lejos de que

---

127 La Ley 23/2006 se publicó el 8 de julio de 2006 en el número 162 del BOE entre las páginas 25561 y 25572 y se puede consultar íntegramente en https://www.boe.es/buscar/doc.php?id=BOE-A-2006-12308 [consulta: 15/06/2023].

128 El proyecto *EuropeanaLocal* se mantuvo activo desde el 1 de junio de 2008 hasta el 31 de mayo de 2011. Durante este periodo, se digitalizaron unos 5 millones de ítems. Se puede encontrar más información en la dirección https://pro.europeana.eu/project/europeanalocal [consulta: 16/06/2023].

129 *eContentplus* abarcó desde el 2005 al 2008. El programa completo se encuentra en la dirección https://eur-lex.europa.eu/ES/legal-content/summary/econtentplus-2005-2008.html [consulta: 16/06/2023].

un ordenador interprete una imagen al mismo nivel que un niño de dos años (Szeliski, 2011, p. 3).

En el campo filológico y, en concreto, en cuanto a los incunables y post-incunables, la visión por computador se ha aplicado en trabajos recientes para discriminar si el impreso utiliza letra gótica o redonda (Seuret *et al.*, 2019), para interpretar el contenido de las ilustraciones (Kim *et al.*, 2020) o para establecer el impresor de ediciones *sine notis* (Lacasta *et al.*, 2022)[130]. Todas ellas buscan en la informática una forma de automatizar las habilidades humanas y, en este sentido, una de las aplicaciones más importantes de la visión artificial es el reconocimiento automático de caracteres (OCR), que permite, en última instancia, que un computador sea capaz de transcribir el texto de una imagen.

Aunque existen diversos tipos de OCR, los que nos interesan por su relación directa con la transcripción literaria son aquellos aplicados a impresos o a manuscritos, denominados específicamente como HTR[131]. En este sentido, las primeras aportaciones basadas en la probabilidad se las debemos a Emanuel Goldberg que, en 1931 presentó, en el Congreso Internacional de Fotografía de Dresden, un artículo en el que desgranaba el diseño de una máquina capaz de indexar un conjunto de microfilms mediante lectura óptica: la *máquina estadística* (Buckland, 1992, pp. 288-289; Goldberg, 1932). Hoy en día, el OCR aplicado a los impresos posteriores al siglo xix obtiene tasas de acierto cercanas al 100% (Breuel *et al.*, 2013, p. 686), por lo que se considera un problema solucionado (Doermann & Tombre, 2014, p. 256); sin embargo, aún se requiere una intervención manual cuando se aplica esta técnica a obras de siglos anteriores (Bazzaco, 2020, p. 545; Weichselbaumer *et al.*, 2020, p. 75).

Por tanto, la digitalización documental es solo el primer paso, dado que no deja de ser una imagen que no permite aprovechar totalmente algunas de las posibilidades de la computación como son la búsqueda y el análisis textual. Ante la ingente cantidad de información que cada día se vuelca en la red, es esencial disponer de mecanismos que nos permitan encontrar los documentos más relevantes. Aunque una imagen se puede acompañar de su descripción bibliográfica y palabras clave, esta información se introduce de manera manual y, por ello, puede resultar incompleta o desviarse de otras descripciones que identifican un mismo producto bibliográfico, lo que es especialmente importante cuando el ejemplar está mútilo, con falta

130 Se averigua el impresor comparando la letra M del impreso con la información contenida en el TW de forma automática.

131 HTR viene de las palabras inglesas *Handwritten Text Recognition*.

de portadas o colofones. Al estar almacenada la digitalización como una colección de imágenes y dado que los buscadores de Internet, en la actualidad, no incorporan OCR, sino que únicamente indexan texto, este *modus operandi* imposibilita encontrar una obra a partir de un fragmento de su contenido. En este sentido, resultaría de interés que un facsímil digital se acompañase de su transcripción, lo que, a su vez, permitiría el volcado a una base de datos, sobre la cual elaborar eventuales búsquedas.

Asimismo, la transcripción automática permite la generación de textos a gran escala en poco tiempo y abre la posibilidad a su tratamiento informático con fines de investigación como la *estilometría computacional* (Daelemans, 2013; Koppel *et al.*, 2009; Stamatatos, 2009), que aprovecha la capacidad que tienen los ordenadores para manejar gran cantidad de información con el fin de efectuar comparaciones para categorizar textos y atribuir autorías. Así lo corroboran diversos estudios en este sentido que ya han dado sus frutos, como el realizado sobre una obra atribuida inicialmente al escritor Miguel Bermúdez y que, mediante la aplicación de esta técnica acompañada de análisis filológicos, aportó pruebas fehacientes que demostraban que su verdadero autor era, en realidad, Lope de Vega (Madroñal & Vega García-Luengos, 2021). Pero existen otras aplicaciones, como así demuestran las líneas de investigación sobre *análisis distante*, que tienen como objetivo buscar patrones durante un periodo o época.

> El análisis distante suele aplicar técnicas computacionales de análisis textual (text mining, procesamiento del lenguaje natural, machine learning, etc.) precisamente por la gran cantidad de texto literario que pretende analizar. Esta es una de sus principales características. De esta manera el análisis distante cubre un área macroanalítica inalcanzable para un análisis manual tradicional, una visión panorámica del hecho literario que sin la tecnología computacional no se podría realizar. La tecnología computacional actúa como un microscopio o un telescopio para la ciencia natural, mostrando aspectos del objeto de análisis en niveles (en este caso "macros") inalcanzables para el ser humano (Navarro Colorado, 2019, p. 52).

No obstante, aunque se ha destacado la estilometría computacional y el análisis distante como grandes beneficiarios de la transcripción de las obras por su inherente necesidad de grandes cantidades de documentos para obtener resultados de calidad, obviamente se puede aprovechar en otros campos. De hecho, la transcripción automática de las copias digitales podría aplicarse a las investigaciones filológicas que tradicionalmente

han utilizado técnicas manuales, como la generación de corpus electrónicos y las ediciones textuales en general, dado que se dispone directamente del texto en el ordenador sin necesidad de introducirlo. La clave está en extender su uso, pero para ello, debe generar confianza entre sus potenciales usuarios con la generación de textos con un bajo porcentaje de errores, un hecho que aún está en vías de desarrollo para impresos anteriores al siglo xviii.

3.2. El almacenamiento y la persistencia de la información

Los procesos de digitalización permiten preservar los elementos físicos, al transformarlos en una secuencia numérica, una copia virtual, que no está sometida al envejecimiento en tanto que carece de materialidad y que, además, permite su copia y distribución utilizando medios informáticos. Cuando se habla de digitalización de un documento o de un libro, se entiende que se ha capturado y codificado para su tratamiento informático, pero es importante diferenciar si se ha hecho mediante una transcripción textual, es decir, introduciendo manualmente el texto, o con la captura del contenedor o soporte. En este último caso, que es el objeto de estudio de este punto, se obtienen imágenes que representan al objeto en ese momento temporal[132], mediante el uso de sensores —escáneres y cámaras— que se conectan a un ordenador. Su representación informática, en realidad, pasa a ser una secuencia de números con un formato establecido, que se empaquetan en lo que se ha venido a llamar de forma generalista *archivo*, y que requiere el uso de un software para ser de nuevo interpretable por el ojo humano.

En el contexto de la digitalización, es habitual definir el libro físico como *libro analógico* y, aunque se podría discutir la falta de precisión del término, por asociarse a magnitudes físicas[133], describe perfectamente el proceso de pérdida de información que experimenta el producto bibliográfico. En realidad, las señales analógicas son aquellas que se representan mediante una función continua en el tiempo, esto es, no sufren saltos. Forman parte del mundo que nos rodea, son los sonidos que percibimos y las imágenes que vemos, pero también aquellas que quedan fuera de nuestros sentidos, como las ondas de radio y la electricidad de casa. Frente a ello, una señal digital únicamente puede adoptar una serie de valores concretos, unos dígitos preestablecidos, técnicamente llamados *valores discretos*. El ejemplo más evidente de dispositivo que utiliza este sistema es el ordenador, que, además de ser digital, únicamente trabaja con dos valores que se representan por convenio como *0* y *1*, y, de ahí, el nombre de *binario*.

---

132 La captura de imágenes en distintas fechas ofrece estadios de los testimonios y de su conservación, lo que resulta de un cierto interés para el análisis y estudio del ejemplar en cuestión.

133 La RAE, en la segunda acepción del adjetivo *analógico*, lo define como: "Dicho de un aparato o de un sistema: Que presenta información, especialmente una medida, mediante una magnitud física continua proporcional al valor de dicha información" y en la tercera: "Que se realiza o transmite por medios analógicos". Las otras acepciones hacen alusión a la analogía.

El ojo humano percibe las imágenes a través de los fotoreceptores que forman la retina, capturando una porción de las ondas electromagnéticas que le llegan, ya que "light is just one form of electromagnetic radiation, and occupies only a small portion of the electromagnetic spectrum" (Snowden *et al.*, 2012, p. 134). Dado que la información que percibimos de nuestro entorno nos llega de forma analógica y el ordenador es digital, para poder almacenarla debe someterse a un proceso de conversión analógico-digital, lo que conlleva una inevitable pérdida de información por el hecho de pasarla a valores discretos[134]. En el caso concreto de la captura de imágenes, se utiliza un sensor fotográfico formado por millones de diminutos receptores sensibles a la luz, habitualmente dispuestos formando un rectángulo, que transforman la señal lumínica recibida en un número. Cuantos más sensores se tenga, mayor será la resolución que se conseguirá y, consecuentemente, la captura final será más precisa[135]. Igualmente, su calidad de fabricación influirá en su sensibilidad ante el espectro cromático, lo que determinará el número de colores que es capaz de reconocer y la fidelidad en la conversión de cada color al valor numérico que lo representa. La imagen final que entregará como salida el dispositivo de captura estará formada por una rejilla de píxeles y un número limitado de colores que no dejará de ser una copia degenerada del objeto físico, dado que habrá sufrido una adaptación a una cuadrícula y una conversión a un número acotado de colores.

El resultado de la digitalización estará formado por uno o varios archivos que se tienen que almacenar. Aunque resulta paradójico, la copia virtual también precisa de un espacio físico para guardarse, con sus correspondientes medidas de conservación y preservación con los que atajar sus propios problemas inherentes. Como consecuencia de este hecho, más allá de la resolución empleada en la captura, una de las primeras decisiones a tomar respecto al almacenamiento es el formato del archivo, es decir, la forma en la que se guardarán los números que describen la imagen, que, a su vez, leerá e interpretará un software de visualización. Afortunadamente, existen estándares que detallan

134 En términos matemáticos, para no perder información al digitalizar una señal, la frecuencia de muestreo debe ser como mínimo el doble de su ancho de banda, según el teorema de Nyquist. No obstante, en el proceso de cuantificación o redondeo posterior se pierde precisión, lo que provoca que la reconstrucción de la señal original no sea fidedigna.

135 En el caso del ojo humano, los encargados de formar los colores son los *conos*, situados en la retina. Una persona sin problemas de visión tendrá tres tipos de conos que se corresponden con las longitudes de onda que forman los colores primarios: rojo, verde y azul. Para más información sobre la visión del color, véase Snowden *et al.*, 2014.

la organización interna de los archivos y hay un consenso generalizado sobre el uso de algunos de ellos[136]. Cada uno de los archivos generados, además, requerirá de un soporte físico en el que escribirlo. Es aquí donde entran los dispositivos de almacenamiento que permitirán que esa copia virtual perdure en el tiempo, a imagen y semejanza del papel con la tinta. Precisamente, las primeras unidades de almacenamiento empleaban cintas de papel y cartulinas en forma de tarjeta para almacenar la información, hasta que, a mediados de los cincuenta, comenzaron a utilizarse las cintas magnéticas, que pasaron a emplear las propiedades de los campos magnéticos para guardar la información. Esta evolución permitió aumentar la capacidad de almacenamiento y la velocidad de lectura y escritura, pero ha supuesto un verdadero problema en el campo de la conservación, dado que son sensibles a fuerzas externas[137], además de que van perdiendo paulatinamente sus propiedades y acaban siendo inservibles.

En 1981 apareció un nuevo soporte para música: el disco compacto o CD, que daba el salto a la informática en 1984. Su principal novedad residía en la inmunidad a los campos de fuerza externos. Ya no utilizaba un sustrato con propiedades magnéticas para almacenar la información, sino un policarbonato plástico junto a una capa de aluminio para conseguir un reflejo lumínico. Para almacenar la información en este soporte, se llevan a cabo deformaciones en la capa plástica que luego se pueden interpretar al proyectar un láser sobre ellas. Este fue el primero de la familia de dispositivos ópticos, que se vio ampliada con el DVD y, posteriormente, por el Blu-ray.

Con el aumento de la velocidad de conexión a Internet y el *streaming*, los dispositivos ópticos están cada vez más relegados y, en el ámbito informático, su uso prácticamente ha desaparecido. De hecho, los fabricantes de ordenadores ya no incorporan este tipo de lectores en sus nuevos productos. Por otra parte, aunque las cintas magnéticas se siguen utilizando en situaciones particulares como mecanismo de copia de seguridad por su gran capacidad de almacenamiento y bajo coste, a nivel doméstico también han quedado obsoletas. Lo mismo ocurre con los discos duros, basados en una tecnología similar, que se siguen utilizando únicamente

---

136 El formato más extendido para almacenamiento de imágenes en alta resolución es el TIFF. Para la creación de documentos formados por varias imágenes es habitual emplear el PDF, que permite disponer las imágenes y el texto en páginas a semejanza de un libro físico. Tanto para un formato como para el otro existe software de libre distribución para los sistemas operativos más extendidos que permite su lectura.

137 Tacón (2008, pp. 48-50) describe con detalle las medidas a adoptar para la correcta conservación de las cintas y discos magnéticos.

cuando se requiere almacenar un volumen grande de datos, dado que los discos de estado sólido, los *SSD*, carentes de componentes móviles y mucho más rápidos, han inundado el mercado. Además, la aparición del almacenamiento en la nube también ha permitido prescindir de dispositivos que se utilizaban años atrás para compartir archivos o realizar copias de seguridad como los *pendrives*.

Esta diversidad de lectores y soportes que ha vivido el almacenamiento informático implica una evidente incompatibilidad entre ellos: un lector de CD-ROM no puede leer un disco magnético. Existe una correspondencia entre soporte y dispositivo lector, por lo que, por ejemplo, para acceder a un archivo que está alojado en una cinta magnética, es necesario disponer de un lector de cintas magnéticas. Esta circunstancia llega a provocar situaciones que no se dan en cuanto a los libros físicos, tal como ocurrió a finales del siglo pasado, cuando

> para conmemorar el 900 aniversario del *Domesday Book* en 1985, se preparó una nueva versión multimedia. En 2002 parecía que el disco ya no era legible, pues escaseaban los ordenadores capaces de leer el formato utilizado. Para salvar la situación se creó un sistema capaz de acceder a los discos mediante técnicas de emulación. Curiosamente, mientras había dificultades para acceder a unos datos digitales de 1986, todavía puede consultarse el *Domesday Book* original, que tiene ya más de 900 años. (Comisión de las Comunidades Europeas, 2005, p. 8)[138]

Para no llegar a que se produzcan este tipo de contratiempos, es importante tomar medidas adecuadas en cuanto a preservación que aseguren una correcta lectura de los datos en un futuro. No solo hay que digitalizar, sino asegurarse de que se dispone de una política de almacenamiento a largo plazo que permita, por un lado, evitar la pérdida del conocimiento y, por el otro, difundirlo.

> La problemática de la preservación conlleva la adopción de soluciones estables de almacenaje en forma de repositorios digitales que deben garantizar la seguridad y la accesibilidad a estos documentos. (Alberch, 2009, p. 135)

---

138 Hoy en día se puede consultar el registro de propiedades que figura en el libro sobre un mapa interactivo en https://opendomesday.org/map/ [consulta: 21/06/2023].

Los repositorios digitales, a nivel de usuario, son un mecanismo de depósito y salvaguarda del material digital. Su funcionamiento es extremadamente sencillo para el usuario: se suben los archivos, se introducen sus metadatos, se establecen los permisos de acceso y, automáticamente, queda accesible en Internet. No hay que preocuparse de migrar los archivos a nuevos soportes ni hacer copias de seguridad. Los repositorios y, de forma general, los servicios de almacenamiento en la nube, se encargan automáticamente de la preservación de los archivos y su difusión, liberando al usuario de esta tarea; sin embargo, no hay que olvidar que, en última instancia, estos archivos se van a almacenar en un soporte físico que, necesariamente, deberá estar sujeto a su política de mantenimiento y prevención de desastres que asegure que la información almacenada perdure en el tiempo. Por ello, cabe plantearse durante cuánto tiempo estarán accesibles. Las empresas privadas pueden cerrar o dejar de ofrecer ese servicio[139] y, con ello, apagar sus servidores con la consiguiente pérdida de información, un hecho que es más raro que ocurra en el caso de repositorios institucionales. Además, ni las grandes empresas descartan que se pueda llegar a producir una hipotética pérdida de datos por falta de mantenimiento, bien porque los recursos económicos y, por ende, los humanos, se vean recortados, o bien por una deficiente política de mantenimiento y gestión de desastres[140]. Para remediar esta situación, se creó el sistema LOCKSS[141], basado en la distribución de copias del contenido en otros repositorios con los que se mantiene una colaboración, y que disminuye la posibilidad de pérdida al almacenar los datos de forma redundante en múltiples localizaciones.

---

139 En septiembre de 2021, Samsung cerró su servicio de almacenamiento en la nube. Se puede ampliar la información en https://www.blomp.com/samsung-closes-down-its-cloud-storage. A finales de 2023, hizo lo propio Amazon con su servicio Amazon Drive, tal como se atestigua en la siguiente dirección https://www.amazon.com/b?ie=UTF8&node=23943055011 [consulta: 20/06/2023].

140 Microsoft ha reconocido que se puede perder información almacenada en la nube. Se puede ampliar la información en esta dirección https://weareproactive.com/cloud-storage-data-loss-is-possible/. De hecho, el servicio en la nube de Amazon sufrió una pérdida de datos en el 2011 que afectó a algunos clientes, a los que contactó por correo electrónico avisándoles de la situación. Se puede consultar una copia de estos correos en https://www.businessinsider.com/amazon-lost-data-2011-4 [consulta: 21/06/2023].

141 El nombre de LOCKSS proviene de las siglas *Lots of Copies Keep Stuff Safe*. Es un sistema desarrollado por el conjunto de librerías de la Universidad de Stanford para asegurar la preservación digital. Se puede encontrar más información en la página del proyecto https://www.lockss.org/ [consulta: 21/06/2023].

Un trabajo de digitalización no se limita a subir los archivos al repositorio, sino que también se deben de etiquetar correctamente, es decir, catalogarlos. De esta forma, las imágenes podrán ser indexadas por los buscadores y encontradas por los potenciales lectores, dado que este tipo de herramientas son textuales. Este proceso se lleva a cabo mediante la definición de lo que se llaman los *metadatos* y, para ello, existen dos mecanismos básicos: una aproximación basada en texto introducido manualmente o mediante *content-based image retrieval* (CBIR) (Lee, 2001, p. 104).

En la aproximación basada en texto, la más común, se define una estructura del catálogo con varios niveles y unos campos descriptivos comunes que forman la ficha bibliográfica que, en el caso de las bibliotecas, están estandarizados con el formato MARC21[142]. Entre los campos que se definen figura el título, autor, impresor, lugar de impresión y fecha. Es una estructura aparentemente sencilla, pero que, además de requerir una labor manual de introducción de datos con sus posibles errores tipográficos, en algunos casos se ve dificultada por la carencia de información objetiva respecto al dato que se graba, dado que no siempre se tiene certeza de sus valores. Esta circunstancia es especialmente llamativa en el caso de obras medievales en las que se puede encontrar el ejemplar carente de portada y/o colofón o como facticio[143], lo que dificulta la labor de clasificarlo correctamente. Esto puede desembocar en tener ejemplares de la misma edición con diversos títulos en distintas bibliotecas lo que, en última instancia, complica la búsqueda posterior por parte de los usuarios; o, al revés, como ocurre con los impresos poéticos incunables iniciados por una misma obra, como las *Coplas de la vita Christi* de fray Íñigo de Mendoza, que se imprimen en ocho ocasiones entre 1482 y 1499. En este sentido, es paradigmático lo ocurrido con dos de sus ejemplares, de los que se pensó que eran una edición diferente de la que, en realidad, se trataba: al catalogar los incunables de bibliotecas estadounidenses, Goff (1973, M-489) identifica la edición del ejemplar de la Library of Congress de Washington como la zaragozana de los Hurus de 1495, cuando, en realidad, se trataba del único ejemplar conservado del incunable salido de las prensas burgalesas de Fadrique de Basilea antes de 1491; de la misma manera, el catálogo de la Bibliothèque de l'École nationale supérieure des Beaux-Arts de París

142 Para una descripción detallada del formato MARC21 se puede consultar la documentación técnica sobre su uso que ha publicado la BNE en https://www.bne.es/es/publicaciones/marc21-registros-bibliograficos [consulta: 23/07/2024].

143 Martos (2023, p. 208) da cuenta de la incompleta e incorrecta catalogación de la reproducción digital del volumen facticio conservado por la Biblioteca Històrica de la Universitat de València con signatura CF/4, conocido como el *Nazareno*.

atribuye su ejemplar a esa misma edición de 1495, cuando, en realidad, pertenecía a la edición de 1492 que, entonces, se consideraba perdida, hasta que Fernández Valladares (2019, p. 54) ha descubierto recientemente la confusión.

El otro mecanismo de introducción de datos, el CBIR, intenta automatizar el proceso de captura de los metadatos. Para ello, se basa en la idea de obtener los valores a partir de la aplicación de un OCR a las imágenes; sin embargo, debido a su escasa precisión en ciertos casos, hace que su uso esté muy limitado y no se aplique habitualmente o se haga una aproximación combinada de ambas técnicas.

Es evidente que un proceso de digitalización no es una tarea sencilla. Hay que elegir la resolución de captura, hacer las tomas con un ambiente adecuado de luz para que las páginas no reflejen y se mantenga la fidelidad de los colores, se debe seleccionar el formato de los archivos, almacenar el material obtenido con una política de preservación a largo plazo y establecer los metadatos. A todo ello, hay que sumarle el trabajo previo de priorización del material a digitalizar y, dando un paso más, qué elementos capturar. De hecho, es habitual quedarse en la mera recolección fotográfica de cada una de las páginas que componen la obra y, aunque como mecanismo de difusión del continente es suficiente, se obvian algunos elementos materiales de los objetos, algunos de los cuales son esenciales para la codicología y la bibliografía material, como es el caso de las filigranas[144], los corondeles y los puntizones, que, al quedar olvidados, obligan, irremediablemente, a la manipulación del libro físico para su estudio.

Una vez obtenida la digitalización del libro, su copia o facsímil digital se puede tratar informáticamente como un archivo o conjunto de archivos más. Se pueden hacer múltiples copias, aplicar algoritmos computacionales, enviarlo como un adjunto de un correo electrónico o hacerlo accesible en Internet a través de un repositorio institucional o una biblioteca digital.

144 Para un estudio de los métodos de reproducción de filigranas véase Díaz de Miranda, 2014.

3.3. Los mecanismos de difusión

La red de redes comenzaba su andadura a principios de los años 70 en un entorno universitario, por lo que sus primeros usuarios, normalmente investigadores pertenecientes a alguna institución, rápidamente vieron el potencial de difusión que les permitía este nuevo medio. El ambiente de colaboración que se vivía con el nuevo sistema propició que se montaran, primero, servidores FTP y, posteriormente, Gopher, antecesores de los actuales repositorios, en los que depositaban artículos para compartirlos con el resto de la comunidad. Con la aparición de la web en los 90 y su acceso libre a la información, este movimiento cobró fuerza, impulsado por las instituciones, que se cuestionaban el costoso y lento sistema de publicación en papel. Auspiciados por este ambiente, un grupo de investigadores de la rama de la física creaba en 1991 *arXiv*, un portal al que subían sus artículos en una fase preliminar para que fueran leídos y comentados por sus compañeros de forma totalmente gratuita. Esta web abrió el camino hacia el *open access* (OA) (Keefer, 2007, pp. 206-207).

Lejos de quedar en una anécdota, en el otoño de 2002, el *Massachussetts Institute of Technology* (MIT) ponía en funcionamiento el primer repositorio institucional[145] y distribuía su código fuente en Internet: el movimiento OA cobraba fuerza. A partir de ese momento, fue una cuestión de tiempo que el resto de instituciones siguieran sus pasos y pusieran en funcionamiento su propio repositorio, extendiéndose este modelo aperturista al resto del mundo (Lynch, 2003, p. 1). En el caso de España, su adopción fue incluso anterior a la del MIT, dado que el primer repositorio que se creó fue *Tesis Doctorals en Xarxa* (TDX) en 2001. No obstante, su verdadera extensión entre las universidades y centros de investigación se produjo a partir del 2005 (Melero, 2008, p. 1). Actualmente, Google Scholar registra 151 repositorios y portales OA[146] en nuestro territorio[147].

Aunque a lo largo de estos años ha ido variando su definición (Lynch, 2003, p. 1; McDowell, 2007, p. 5; Abadal, 2012, p. 23), de forma generalizada, un repositorio institucional es

145 El software de repositorio que creó el MIT se llama *DSpace* y se desarrolló en colaboración con la empresa informática Hewlett Packard (Lynch, 2003, p. 1).

146 La información se ha extraído del informe que prepara anualmente el CSIC a través de la web *Ranking Web of Repositories*.

147 Google requiere que los metadatos de la página cumplan unas características para indexarla como repositorio *open access*, por lo que seguramente este número es mayor.

> un conjunto de servicios prestados por las universidades y organismos de investigación, al conjunto de la comunidad, para recopilar, administrar, difundir y preservar la producción documental digital generada en la institución, cualquiera que sea su tipología, a través de la creación de una colección digital organizada, abierta e interoperable a través del protocolo OAI-PMH (protocolo para la recolección de metadatos) con el fin de garantizar un aumento de la visibilidad e impacto de la misma. (Ferreras Fernández, 2018, p. 44)

Por lo tanto, la tipología de archivos no está limitada, pudiendo albergar tesis, artículos, revistas, libros electrónicos, presentaciones y grabaciones, entre otros. Aunque el número de repositorios que existen actualmente es cuantioso, suelen compartir una interfaz común, dado que el uso del software del MIT como base para su creación está muy extendido. Su característica básica es la sencillez, presentando un buscador junto a unas categorías y, ocupando la mayor parte de la interfaz, el listado de archivos.

El otro gran difusor del conocimiento en la red y hábitat natural de las digitalizaciones de obras literarias son las bibliotecas digitales. Nacieron casi al mismo tiempo que Internet y, así, Michael Hart iniciaba en 1971 el proyecto Gutenberg como un sistema rudimentario de distribución de textos electrónicos que había transcrito previamente de forma manual[148]. Aunque su intención era que todo el mundo tuviese acceso a la información, en realidad, había inventado el *eBook* y la biblioteca digital sin percatarse de ello. Actualmente, este portal alberga más de 70 000 referencias de libre distribución que cientos de voluntarios han ayudado a transcribir para mantener su enfoque fundacional de difusión del material en formato textual[149].

La definición de qué es una biblioteca digital, al igual que ha ocurrido con la de repositorio digital, ha ido evolucionando y perfilándose con los años, pero de forma sencilla:

---

148 El primer documento que transcribió fue la Declaración de Independencia de los Estados Unidos, según apunta la web del proyecto Gutenberg en la página https://www.gutenberg.org/about/background/50years.html [consulta: 23/09/2023].

149 No solo hacen transcripciones de obras puramente textuales, sino también de tratados científicos, como *A Course of Pure Mathematics* de G. H. Hardy, que suponen un gran trabajo en la reescritura de las fórmulas matemáticas.

> digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible. (Lesk, 1997, p. XIX).

Aunque esta es una de las primeras definiciones, no por ello ha perdido vigencia. En principio, se podría aplicar también a un repositorio por su visión generalista de lo que es una colección digital; el matiz que marca la diferencia radica en la relación con el material que almacenan las bibliotecas. Es decir, pese a su similitud, un repositorio institucional alberga información creada por la comunidad que lo gestiona, al contrario que la biblioteca digital, que se estructura y recupera la información a imagen y semejanza de una biblioteca física, lo que conlleva que su tipología documental sea más limitada. De hecho, como se verá más adelante, existen ejemplos de convivencia de ambos en los que los objetos de la biblioteca digital residen en el repositorio de la institución que gestiona ambos. Con su creación:

> la bibliothèque, qui fonctionnait selon une logique de stock (conserver les imprimés et autres supports), doit passer pour partie à une logique de flux, soit en reversant dans les nouveaux supports les contenus préexistants (c'est le principe de la digitalisation et des bibliothèques virtuelles), soit en créant directement en ligne des contenus eux-mêmes nouveaux. (Barbier, 2013, p. 287)

Una biblioteca digital bien puede estar especializada en una temática o época, o bien ser generalista y albergar, así, una amplia variedad de recursos como obras textuales, fotografías, mapas, sonidos, música e imágenes en 3D. Las obras textuales pueden encontrarse transcritas, ser un facsímil digital o una combinación de ambas. Desde la creación de Internet y la aparición del proyecto Gutenberg, continuamente aparecen nuevos portales que proporcionan contenido digital.

Entre las bibliotecas digitales de índole generalista que albergan incunables, una de las primeras que apareció en España fue la Biblioteca Virtual Miguel de Cervantes (BVMC). Se publicó en 1999 y, al igual que el proyecto Gutenberg, nacía sin estar ligada a una biblioteca física concreta. Asimismo, desde un primer momento, se acogía a una política OA, distribuyendo el material sin necesidad de suscripción y de forma totalmente gratuita. Está construida como un metaportal o portal de portales en el que se estructura el contenido en áreas temáticas que comparten una interfaz

común. Uno de estos portales, como ya se ha apuntado anteriormente, es la Biblioteca Joan Lluís Vives, un espacio de difusión de la literatura catalana que abarca desde el Medievo hasta la actualidad, con secciones dedicadas en exclusiva a la vida y obra de Ausiàs March, Ramon Llull o Joan Roís de Corella, así como un primer intento de portal temático de la asociación *Convivio*, centrada en la poesía medieval, sin actualización desde 2012.

Es evidente que, pese a su magnífica organización, la cantidad de información disponible puede resultar abrumadora para el neófito, por lo que, si se busca una obra en concreto, lo más rápido es acudir al buscador, que permite la acotación de la información mediante diversos parámetros e, incluso, discernir entre la búsqueda únicamente en catálogo o en el texto de las obras. Es de destacar este último caso, ya que acompañará a cada uno de los ítems con un fragmento del contenido de la coincidencia a modo de concordancia.

La página de resultado de la búsqueda muestra un listado con las fichas de las obras que coinciden con las palabras clave, así como un panel lateral con el número de elementos según el tipo de obra, formato y autor, entre otros y que, a su vez, permite delimitar aún más la búsqueda. Respecto a las fichas bibliográficas, cada una tiene sus campos enlazados, por lo que se crea una red de relaciones que permite navegar entre autores, fechas, materias e instituciones donde se alberga el objeto físico o digital, en un magnífico ejemplo de explotación de las posibilidades que nos brindan el hipertexto y la web semántica.

Dada la heterogeneidad de portales que alberga la BVMC, y aunque todos ellos conservan la misma interfaz, cada uno se especializa en un contenido y, por ello, dispone de unas secciones internas propias según el área en la que estén catalogados. Por ejemplo, las bibliotecas de autores, dentro del área de literatura, tienen secciones con su biografía y su obra, a diferencia de las instituciones que, al no tener sentido dicha funcionalidad, carecen de ellas.

Las obras literarias se encuentran o bien transcritas o bien como facsímiles digitales. En este último caso, se dispone de una colección de unos 50 000 ejemplares que se pueden consultar en una sección dedicada dentro del área de literatura. Estas digitalizaciones están en formato PDF o como un conjunto de imágenes independientes a las que se accede a través de una página HTML que contiene el índice. La diferencia entre ambas, además del formato de almacenamiento, es que a los PDF se les ha aplicado un reconocimiento de caracteres automático que permite la selección del texto; de hecho, resulta curioso que, de momento, su buscador interno no indexe ese contenido, por lo que no será posible encontrar dichos documentos mediante palabras clave contenidas en ellos.

Por su parte, la interfaz de exploración de las imágenes de los facsímiles en HTML es sencilla y carece de funcionalidades extras, como visualización de miniaturas, exportación a PDF u organización de los capítulos mediante paneles laterales con enlaces. La página de entrada a la digitalización ya deja entrever la sobriedad, al mostrar un listado en bruto de los enlaces a cada una de las imágenes de las páginas digitalizadas. A su vez, la interfaz de navegación a través del facsímil, actualmente, tampoco presta servicios adicionales más allá de los enlaces indispensables que permiten pasar de página y volver al índice, como podrían ser la posibilidad de exportar una selección de páginas a PDF, ampliar o reducir la imagen o llevar a cabo anotaciones sobre ella.

Un año después de crearse la BVMC, comenzaba el proceso de generación de la colección *Somni* a partir de la digitalización de los fondos la Biblioteca Històrica de la Universitat de València, poseedora de un rico patrimonio de gran interés filológico con un total de 356 incunables.

> La digitalización de los fondos se hacía a partir de su soporte en microfilm, generándose imágenes digitalizadas con una resolución Tiff compresión Fax grupo IV en blanco y negro (.gif), por ser un formato con el que se ocupaba un menor espacio, a la vez que ofrecía una buena visualización. (Millás Mascarós & Escriche Soriano, 2017, p. 5).

No obstante, tal como se ha apuntado anteriormente, a partir del 2010, a raíz de su participación en el proyecto *Europeana Regia*, se sustituyó la microfilmación por la digitalización directa de los originales, con la consiguiente obtención de imágenes de alta resolución y, a su vez, para disponer de preservación digital, se integró la colección en el repositorio institucional de la universidad (RODERIC).

En relación a la funcionalidad, desde la página principal de *Somni* se puede acceder directamente a sus colecciones, entre las que figuran las de los manuscritos del Duque de Calabria, las de mapas, las de grabados y las de incunables, entre otras. Al entrar a cada una de ellas, se obtiene un listado con los principales campos de las fichas bibliográficas acompañados de una miniatura de la portada. El menú de la parte lateral permite acotar los resultados por autor, materia y fecha de publicación. Esta página es la misma que se obtiene en caso de que se utilice el buscador que, además, da la posibilidad de abarcar todo el repositorio o únicamente la colección *Somni*.

En lo que respecta a los incunables, la catalogación completa contiene la descripción general, la descripción física y un conjunto de referencias a

otros catálogos. En ningún caso, estos campos son hiperenlaces que permitan, por ejemplo, navegar entre registros del mismo autor o del mismo impresor. Por su parte, el acceso a las imágenes se lleva a cabo mediante una interfaz que, además de la funcionalidad básica que permite pasar de una página, contempla la posibilidad de exportar a PDF, rotar las imágenes y hacer zoom en ellas con la rueda del ratón.

En 2002, poco después de comenzar la creación de *Somni*, la Generalitat Valenciana ponía en marcha la Biblioteca Valenciana Digital (BIVALDI), un portal para dar difusión a los fondos de la Biblioteca Valenciana Nicolau Primitiu (BV). Actualmente, aloja más de 9 000 obras digitalizadas, de las cuales 5 000 son libros[150], ejerce de agregador de *Hispana* y *Europeana* y comparte sus registros con la Biblioteca Virtual de Patrimonio Bibliográfico y la World Digital Library. Aunque en su mayoría son obras que pertenecen al fondo de la BV, también contiene algunas digitalizaciones que resultan relevantes en el patrimonio bibliográfico valenciano, como es el caso del único folio conservado del incunable de la *Bíblia Valenciana* de 1478, cuyo original se conserva en la *Hispanic Society* de Nueva York, o del incunable de *Les trobes en llaors de la Verge Maria*, impreso en Valencia por Palmart en 1474, que se encuentra en la Biblioteca Històrica de la Universitat de València.

La página principal de BIVALDI permite acceder rápidamente a la consulta de su catálogo, a la sección de hemeroteca, perfectamente integrada, así como a las diversas colecciones, entre las que figuran los incunables. Ya dentro de las colecciones, se muestra un listado en el que destacan las imágenes reducidas de una página de la digitalización, acompañadas en la parte inferior del título y del autor. Esta visualización confirma la relevancia que han querido otorgarle al facsímil digital por encima de su estudio bibliográfico. De hecho, se accede directamente a su visualización si se pincha en la miniatura. En caso de querer ver su ficha bibliográfica, habrá que hacer clic sobre el título.

Cada una de las fichas de BIVALDI, en consonancia con las de la BVMC, vuelve a ser una muestra de las posibilidades del hipertexto, creando una red de relaciones a través de campos enlazados que permiten la navegación a otras fichas bibliográficas del catálogo del mismo autor, autores secundarios o registros relacionados. Asimismo, incluye la localización de cada uno de los ejemplares y, según la obra, una colección de recursos como la transcripción, bibliografía, la propia digitalización e imágenes

150 Se pueden consultar las estadísticas actualizadas en la página web de la biblioteca.

técnicas relacionadas con dicho proceso. Por supuesto, en un portal con estos alardes técnicos no podía faltar la posibilidad de exportar los registros del catálogo a multitud de formatos, como MARCXML, BibTex, Dublin Core RDF y Link Open Data/EDM 5.2.8, entre otros.

La digitalización de la obra se muestra en una capa superpuesta que contiene una tabla con las imágenes en miniatura de cada una de las páginas, y que sigue el modelo estético del listado del catálogo. Dispone de un diseño muy intuitivo, que facilita en gran medida la navegación a través de las páginas y que proporciona diversas herramientas de apoyo, con una disposición y una iconografía que permiten al usuario sentirse cómodo desde el primer momento. Es el caso de la barra lateral que, de forma permanente, sugiere enlaces a la bibliografía relacionada, a los derechos de autor y, en algunas ocasiones, a la transcripción completa. La experiencia de navegación a través de la obra digitalizada se acompaña con una respuesta rápida al cambio de página o al zoom, aunque en este último caso, hubiese sido deseable un mayor aumento. La lista de opciones se completa con la posibilidad de exportar las imágenes tanto en formato JPG como en PDF.

Dada la prontitud con la que aparecieron las anteriores bibliotecas, sorprende que la Biblioteca Nacional de España (BNE) dilatase la puesta en marcha de su particular proyecto, la *Biblioteca Digital Hispánica* (BDH), hasta 2008, un año después de abrir la Hemeroteca Digital. Pese a ello, apareció con un nutrido catálogo compuesto por 10 000 obras que ha ido engrosando a lo largo de los años, contando con más de 200 000 referencias en el 2023.

El diseño de la interfaz de la BDH hace destacar el buscador como herramienta principal de acceso a su vasta colección. Por supuesto, permite la búsqueda avanzada con el correspondiente filtrado de tipología documental, así como en el interior de los textos de las digitalizaciones —opción que no permite la BVMC—, dado que también se les ha aplicado un reconocimiento de caracteres automático básico integrado en el software de captura, con una buena precisión en impresos contemporáneos. Pero, si lo que desea el usuario es explorar el contenido en toda su amplitud, también ofrece las típicas colecciones, entre las que figura, por supuesto, la de incunables españoles.

El resultado de las búsquedas y de las colecciones está compuesto por las consabidas miniaturas, acompañadas del título, autor, formato y año de publicación, junto a una barra lateral para filtrar las obras. No obstante, a diferencia del resto, permite la posibilidad de cambiar el formato de la retícula para otorgar mayor o menor importancia a las miniaturas, que, desgraciadamente, tienden a limitarse a las cubiertas de las obras. Como funcionalidad complementaria, el pie de cada registro se acompaña con

unos botones iconográficos para obtener el enlace permanente de la ficha o de la digitalización, publicar en redes sociales o dejar un comentario.

La digitalización se muestra con una interfaz que dispone de las opciones de navegación habituales, así como de rotación, centrado, impresión y descarga. Aunque el cambio de página no es rápido, posiblemente sea debido a la calidad con la que se descargan las capturas, algo que se puede apreciar cuando se hace uso del zoom que permite una ampliación de hasta el 500% sin que se aprecie una pérdida de detalle, salvo excepciones.

La Biblioteca de la Universidad Complutense de Madrid, la cuarta de España en cuanto a número de incunables, empezó el proceso de digitalización de sus fondos en el año 1995 con el llamado proyecto *Dioscórides*, que incluía la colección biomédica. Desde entonces, se ha ido ampliando con sucesivos acuerdos, colaboraciones y ayudas de instituciones públicas y privadas[151]. Actualmente, bajo el nombre de *Patrimonio Digital Complutense* (PDC), alberga diversas colecciones, entre las que se encuentran los manuscritos, incunables, impresos antiguos, grabados, mapas, fotografías y tesis doctorales históricas.

El acceso a los materiales comparte una interfaz homogénea acorde a los estándares actuales con un diseño *responsive*. El listado de objetos de cada colección se muestra con una miniatura de la primera página relevante junto a su título, autor y fecha de impresión. La ficha interna da acceso, mediante botones iconográficos, a los metadatos que, en el caso de los incunables, incorporan un campo de notas donde se hace una breve descripción material y otra información relevante, como la fecha de ingreso en la biblioteca. El visualizado de la digitalización se realiza con una herramienta que se muestra como una capa flotante superpuesta que ocupa un porcentaje de la ventana principal del navegador, sin posibilidad de colocarla en una ventana independiente ni ampliarla a pantalla completa, lo que dificulta la visualización de las imágenes.

Desde el 2014 se puede acceder a través de la *Biblioteca Patrimonial Digital* (BiPaDi) a una selección del fondo antiguo de la Biblioteca de la Universitat de Barcelona. Dispone de un gran número de colecciones,

151 El Proyecto *Dioscórides* se desarrolló a partir de un convenio de colaboración entre la Universidad Complutense de Madrid y los laboratorios Blaxo-Wellcome. Se puede obtener una relación de los distintos proyectos de digitalización para formar la biblioteca digital que se han ido sucediendo desde sus inicios hasta la actualidad en https://patrimoniodigital.ucm.es/s/patrimonio/page/proyectos [consulta: 15/10/2023].

entre las que se encuentra la de incunables[152], que se muestran como un listado vertical de miniaturas, con diseño *responsive*, y que se acompañan del título, el autor y la fecha de impresión, así como la habitual barra lateral que facilita la acotación de los resultados.

La ficha de cada una de las obras enumera las características más relevantes, aunque no se incluye ningún apartado de materialidad más allá de la especificación del formato. La visualización de las capturas se realiza con una herramienta con las opciones básicas para pasar página, rotar y hacer zoom, pero que tiene la opción de mostrar cada una de las imágenes a pantalla completa, lo que facilita su estudio y, aunque sencilla, cumple a la perfección con su propósito. De hecho, es la misma que utiliza la Biblioteca de Catalunya para la exhibición de sus casi 600 incunables que atesora entre sus fondos.

Más allá del contexto hispánico, en 1997 apareció *Gallica*, la versión digital de la Bibliothèque nationale de France (BnF), con imágenes y textos principalmente decimonónicos (Bordier, 2023, p. 131):

> Les contenus de Gallica correspondent initialement à une bibliothèque encyclopédique de culture francophone destinée aux travaux d'enseignement et d'érudition, mais aussi devant permettre la découverte de ressources culturelles par le grand public. (Lupovici *et al.*, 2003, p. 41)

En la actualidad, contiene casi 10 millones de documentos que comprenden libros digitalizados, periódicos, cartas, fotos, pinturas y otros menos comunes como mapas, videos, partituras y audios. Afortunadamente, esta vasta colección viene acompañada de un potente buscador que permite acotar los resultados por múltiples parámetros, como el tipo de documento, la temática o la región geográfica. El acceso también se puede llevar a cabo a través de colecciones. Aunque no hay una dedicada exclusivamente a los incunables[153], existe un camino si se accede primeramente al conjunto completo de impresos, que presenta el listado de registros acompañados

---

152 La colección de incunables reside en el *Centre de Recursos per a l'Aprenentatge i la Investigación* (CRAI) de la Universitat de Barcelona, en su totalidad procedente de conventos suprimidos por la desamortización de Mendizábal de 1835, según se recoge en https://bipadi.ub.edu/digital/collection/incunables [consulta: 28/10/2023].

153 *Gallica* tiene selecciones de obras relevantes de distintas temáticas; entre ellas, existe una dedicada a los orígenes de la imprenta.

de la miniatura de la digitalización. Desde esta pantalla, mediante la barra lateral que acota los resultados, es posible conseguir únicamente las fichas de las obras del siglo xv, aunque la fiabilidad de la catalogación del año es dudosa, dado que, incluso, llega a proporcionar impresos con fecha de datación anterior al xiv.

Gallica no solo muestra los fondos de la BnF, sino también de otras bibliotecas como *Numelyo*, la biblioteca digital de Lyon, *Tolosana*, de la Universidad de Toulouse y *e-rara.ch*, perteneciente a la agrupación de bibliotecas suizas. La integración de los resultados en el mismo listado es perfecta, ya que únicamente se diferencian los registros mediante un pequeño icono situado en el lateral; sin embargo, la visualización del material se lleva a cabo a través de la página de la biblioteca de origen. Esta característica provoca que exista una disparidad de interfaces según donde se encuentre la obra que se consulte, cada una con su propio diseño visual y funcionalidades. Aunque sus diferencias no son notables, destaca la herramienta de visualización de los fondos de la BnF por ser la más completa, acompañada con una iconografía y distribución de elementos en la pantalla que facilitan la consulta. La respuesta al cambio de página es rápida e incorpora el marco de trabajo IIIF[154], que facilita la creación de anotaciones sobre las imágenes, compartir las capturas y ajustar diferentes características, como el brillo y el contraste. En contraposición, la interfaz de *Numelyo*, pese a ser muy completa, acusa lentitud en el cambio de páginas, carece de zoom, de anotaciones y no tiene la posibilidad de ajustar los parámetros de las imágenes.

El mismo año que aparecía *Gallica* también lo hacía la biblioteca digital de la Bayerische Staatsbibliothek (BSB), delegando la digitalización en el Munich DigitiZation Center (MBZ). Como novedad, no solo llevan a cabo la digitalización habitual de material impreso, sino que también capturan modelos en 3D en color que se pueden consultar a través de una herramienta —desarrollada específicamente por el equipo del MBZ— que

154 IIIF es un marco de trabajo que define la forma de almacenar y transmitir imágenes, sonidos y videos que conforman un volumen. Establece la relación entre los elementos que lo componen, por ejemplo, el orden de las páginas y, además, facilita la búsqueda y visualización. Dado que facilita el intercambio de información de objetos digitales, cada vez más bibliotecas lo incorporan, como es el caso de *Gallica*, *Europeana*, *Bayerische StaatsBibliothek* y la *Digital Bodleian*. Se puede ampliar la información en la página web https://iiif.io/ [consulta: 25/11/2023]. El portal *Biblissima* es ejemplo perfecto de uso, dado que permite la búsqueda y consulta en todas ellas desde una misma interfaz, sin necesidad de navegar a la biblioteca donde reside el archivo, que se recupera digitalmente para mostrarlo en su herramienta de visualización.

permite manipular fácilmente el objeto en un entorno virtual y observarlo desde todos los ángulos.

La BSB posee una de las mayores colecciones de incunables, de los cuales unos 8 000 se encuentran actualmente digitalizados y disponibles en línea, con la posibilidad de acceder a ellos mediante IIIF. Dada la cantidad de objetos que almacena, en su diseño visual destaca el buscador como el elemento principal de acceso. Dispone, por supuesto, de los parámetros habituales de filtrado a los que añade la posibilidad de búsqueda en los textos de las imágenes digitalizadas. El resultado, en este caso, muestra un recorte de la imagen con el texto buscado resaltado. Es una funcionalidad que brilla por su extraordinaria ejecución y funcionamiento, dada la complejidad técnica subyacente que implica. Obviamente, esta muestra del dominio de la programación acompaña al resto del portal. Con un diseño *responsive*, el listado de las obras junto a la miniatura se acompaña con un botón que despliega toda la información bibliográfica con algunos de sus campos enlazados, en concreto, los autores y la ficha correspondiente en el GW. Asimismo, incluye información relevante como los diferentes volúmenes que forman una obra o, en el caso de facticios, aporta las referencias a los diferentes materiales reunidos bajo una misma encuadernación.

La disposición de los elementos que conforman la interfaz de digitalizaciones facilita su uso. En términos de funcionalidad, proporciona un zoom considerable sin pérdida de detalle, se pueden realizar recortes, modificar el brillo, el contraste, la saturación, rotaciones, exportaciones a otros formatos para consultarlo en local y compartir en redes sociales. La única opción que no contempla es la posibilidad de realizar anotaciones sobre las páginas, pero si realmente se necesita, se puede utilizar una herramienta de visualización externa, dado que permite la exportación de las imágenes con IIIF.

El volumen de fondos digitalizados por la BSB alcanza el 99%, algo que contrasta con el ritmo de otras bibliotecas. Es el caso de la *Bodleian Digital* de la Universidad de Oxford, otra de las bibliotecas que cuenta con un extenso fondo de incunables, pero que únicamente ha digitalizado en torno al 10%, es decir, no llega a 700 ítems. Pese a ello, presenta una interfaz de consulta *responsive* con un diseño visual limpio sin elementos adicionales que desvirtúen el propósito principal de una librería digital: la consulta de las colecciones. La herramienta de visualización no destaca respecto al resto, es sencilla y rápida, cumple a la perfección su cometido con la funcionalidad básica acompañada de un buen zoom, pero incorpora un elemento sumamente relevante para determinadas áreas de investigación: las capturas se han hecho junto a una regla de medida que permite, con relativa precisión, determinar el tamaño de los elementos que residen en la imagen, como el interlineado o los tipos. Es un objeto sencillo de incorporar que facilita a los

investigadores de la materialidad de la obra la toma de datos sin necesidad de consultar el original.

Este recorrido por algunas de las principales bibliotecas digitales es una muestra representativa de las distintas implementaciones. Aunque todas ellas se basan en un buscador y acompañan los resultados con una barra lateral que los permite acotar, las diferencias se observan en el contenido de las fichas bibliográficas y en la herramienta de visualización de las digitalizaciones. Entre ellas, sobresale la BSB, con unas fichas que destacan por la gran cantidad de información que proporcionan, se encuentran enlazadas a catálogos en los que se amplían los datos y se enmarcan con un diseño de la interfaz de consulta integrado visualmente en la página. La herramienta de visualización dispone de todas las funcionalidades habitualmente requeridas, de uso sencillo y con una interfaz intuitiva. Es, por descontado, un modelo de biblioteca digital a imitar, tanto a nivel de usabilidad como de funcionalidad, y en el que, además, prácticamente se ha finalizado el proceso de digitalización de todos sus fondos.

No cabe duda de que las bibliotecas digitales seguirán creciendo. Se han estandarizado las técnicas de captura de imágenes, con lo que se ha conseguido que sean un fiel reflejo de la realidad, con una alta resolución que permite observar hasta los mínimos detalles, pero, pese a ello, aún queda margen de mejora. En contadas ocasiones, aunque cada vez va siendo más común, se incorporan la carta de color o elementos de referencia para establecer medidas. Sin embargo, no se da un paso más en la accesibilidad —y, en última instancia, en la preservación misma ante eventuales pérdidas del original—, dado que, en ningún caso, se emplean dispositivos que permitan la captura de la *verjura* con sus filigranas, corondeles y puntizones, elementos que ayudarían a su estudio material sin necesidad de desplazarse al fondo que conserva el original para su consulta directa.

4. Inteligencia artificial aplicada a la digitalización de textos

Desde la antigüedad, la replicación del funcionamiento del cerebro humano ha sido un tema recurrente entre los grandes pensadores. El primero del que se tiene constancia es Aristóteles, que elaboró un sistema de silogismos que permitía generar conclusiones a partir de unas premisas. En la Edad Media, Ramon Llull esbozó la teoría de un método con capacidad de razonamiento que combinaba la teología y la filosofía. Más pragmático fue Leonardo da Vinci, que diseñó una calculadora mecánica formada por engranajes y palancas. En 1623, el científico alemán Wilhem Shickard, esbozó en una misiva dirigida a un amigo un artilugio similar que tampoco se llegó a construir. Fue Blaise Pascal, pocos años después, quien finalmente construyó la primera máquina de calcular, con un tamaño similar a una caja de zapatos, que permitía una rápida conversión entre divisas (Russell & Norvig, 2010, pp. 5-6).

Pese a que la automatización de los cálculos matemáticos se remonta al siglo XVII, el nacimiento de la inteligencia artificial (IA) es más reciente. Concretamente, se le atribuye a Warren McCulloch y Walter Pitts (1943) por el modelo teórico que propusieron en el que combinaron las especialidades científicas de ambos: la neurociencia y la lógica. Tomaron como base una representación simplificada de una neurona que únicamente aceptaba dos estados: *encendido* y *apagado*. En consonancia con el modelo biológico[155], la neurona artificial se activaba si recibía suficientes estímulos de las otras neuronas con las que estaba conectada. Con esta estructura en forma de malla, demostraron que era posible calcular funciones y representar

155 El trabajo de McCulloch y Pitts estaba basado en los estudios previos sobre las neuronas y sus interconexiones de Santiago Ramón y Cajal a quien, en 1906, se le otorgó el Premio Nobel en Fisiología y Medicina en 1906 junto a Camillo Golgi por sus trabajos sobre la estructura del sistema nervioso (Nilsson, 2010, p. 34).

operadores lógicos e, incluso, sugirieron que podría llegar a aprender. Esta hipótesis se basa en el hecho de que el mundo que nos rodea se rige por funciones matemáticas. Desde el crecimiento de las plantas hasta nuestra visión, todo es representable mediante una función más o menos compleja, que se puede ver como una caja a la que se le introducen unos valores y arroja como salida otros. El problema reside en que, en ocasiones, es extremadamente difícil modelarlas.

Crear un algoritmo que ordene un conjunto de números de menor a mayor es sencillo y, de hecho, existen múltiples maneras de hacerlo de forma más o menos eficiente; sin embargo, modelar ciertas situaciones, como el comportamiento de un usuario en base a los productos que tiene en su cesta de la compra o la identificación de los objetos de una fotografía, aunque para un humano sea muy intuitivo, es sumamente complejo de describir para que una máquina lo efectúe correctamente. De hecho, los test de verificación web más conocidos, los *captchas*, muestran varias imágenes en pantalla para que el usuario haga clic únicamente en aquellas en las que aparece un determinado elemento, como un semáforo, con el fin de diferenciar si quien accede es un humano o un software. Con la IA y, en concreto, con el aprendizaje automático, se busca que la máquina consiga crear esas funciones, tan enrevesadas de programar, de forma automática.

4.1. Inteligencia artificial y aprendizaje automático

El término *inteligencia artificial*, entendido como disciplina científica, surgió en el seno del *Dartmouth Summer Project*, una reunión que se celebró en Hannover, New Hampshire, en 1956 y que contó con la asistencia de diversos investigadores en este campo. Aunque, al principio, hubo cierta controversia por las posibles connotaciones de falsedad que podía provocar la palabra *artificial*, finalmente se decidió que "it is a good name. Like all names of scientific fields, it will grow to become exactly what its field comes to mean" (Nilsson, 2010, p. 79). Quien pronunció estas palabras fue Allen Newell, que en ese momento estaba trabajando junto con Herbert Simon y Cliff Shaw en un programa que demostraba teoremas matemáticos utilizando lógica simbólica, el *Logic Theorist*, considerado el primer software en exhibir comportamiento inteligente[156]. De hecho, a Newell y Simon se les concedió el prestigioso Premio Turing en 1975 y, en su discurso:

> formularon la hipótesis del sistema de símbolos físicos (SSF), según la cual "todo sistema de símbolos físicos posee los medios necesarios y suficientes para llevar a cabo acciones inteligentes". Por otra parte, dado que los seres humanos somos capaces de mostrar conductas inteligentes en el sentido general, entonces, de acuerdo con la hipótesis, nosotros también somos sistemas de símbolos físicos. (López de Mántaras Badia & Meseguer González, 2017, p. 8)

La hipótesis del SSF apoya el modelo clásico utilizado por los primeros pensadores, la *IA simbólica*[157], que toma como base la lógica[158] para crear representaciones abstractas del mundo real, derivar nueva información y tomar decisiones. Por ejemplo, a partir de las proposiciones

> 1: Un humano es un mamífero
> 2: Juan es un humano,

se deduce, por lógica de primer orden, que Juan es un mamífero.

---

156 *Logic Theorist* fue capaz de demostrar 38 de los primeros 52 teoremas del capítulo 2 de *Principia Mathematica* (Clevier, 1992, p. 46), un conjunto de libros de principios del siglo xx que aglutinaba el conocimiento matemático de la época.

157 El modelo simbólico ha sido el utilizado por los filósofos desde la antigüedad para modelar el razonamiento humano, por lo que también se denomina *GOFAI*, cuyas siglas provienen de *Good Old Fashioned AI*.

158 Los modelos clásicos más conocidos de lógica son la proposicional y la de predicados.

La SSF, en realidad, también se puede aplicar a la otra gran corriente de la IA, la *conexionista*. En este caso, se fundamenta en la interacción de una gran cantidad de elementos que, actuando como un todo, conectados, exhiben un comportamiento inteligente. Su mayor representante es una red neuronal artificial y su forma más simple es el *perceptrón*, inspirado en el modelo de McCulloch y Pitts y formado por una capa de entrada y un nodo de salida.

Una neurona es una célula nerviosa formada por un cuerpo celular con un axón y unas dendritas, tal como se muestra de forma esquemática en la Figura 1. El cerebro está formado por millones de ellas y, aunque no están unidas físicamente, se pueden comunicar por medio de neurotransmisores, unas sustancias químicas que se producen en las terminaciones del axón y que se vierten en el espacio que las separa, la *sinapsis*. La activación de una neurona se produce cuando el número de neurotransmisores recibidos a través de las dendritas supera un umbral, momento en el cual comenzará a estimular, a su vez, a las neuronas vecinas[159]. El impulso nervioso, de esta forma, se irá propagando de una célula a otra formando un camino a través de la red neuronal que desembocará en la activación final de unas neuronas determinadas. El ajuste de la cantidad de sustancia química vertida en el espacio sináptico[160] permitirá modificar el recorrido y, con ello, variar el destino final de la información. Es el proceso conocido como *aprendizaje* y que produce, como resultado, la reconfiguración de la actividad neuronal. Este modelo de funcionamiento del sistema nervioso fue descrito por Ramón y Cajal a finales del siglo xix y sigue totalmente vigente (Lafarga Coscojuela, 1994, pp. 8-12)[161].

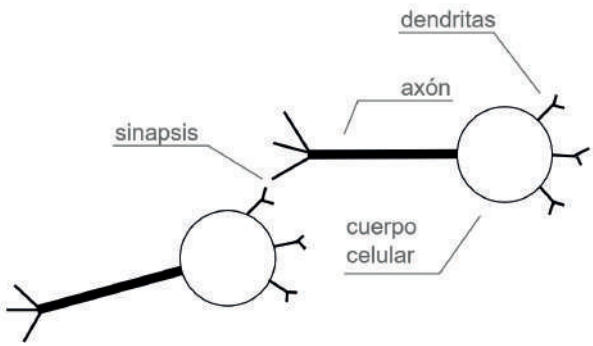
---

159 "La excitabilidad se puede definir como la capacidad de una célula para responder a un estímulo. La respuesta de la neurona va a depender de las características que presente su membrana plasmática. De manera global se puede decir que las neuronas tienen un potencial de membrana en reposo y un conjunto de elementos o señales posibles que definen su propiedad de excitabilidad. Tales elementos para la excitabilidad pueden ser de recepción del estímulo, la integración de los diferentes estímulos para generar un potencial de acción, la conducción de este potencial hacia los sitios efectores de la neurona (terminales sinápticas) y la transmisión del mensaje a otra neurona" (Quintanar, 2010, p. 35).

160 Lugar en el que se produce la sinapsis, que la RAE define como: "conexión entre el axón de una neurona y la dendrita de otra cercana mediante neurotransmisores".

161 El cerebro es un órgano cuyo funcionamiento intriga a los investigadores y en el que aún se están descubriendo nuevos elementos que lo forman. Aparte de las neuronas, está compuesto por células glías, que actúan como soporte, así como por astrocitos glutamatérgicos, que son un híbrido entre ambas (De Ceglia *et al.*, 2023).

Figura 1. Neurona biológica



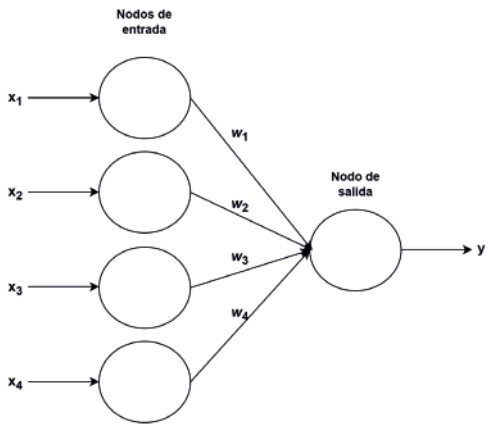
El perceptrón, que se muestra en la Figura 2, es la unidad computacional de una red neuronal artificial. Se inspira en el modelo biológico de una neurona para simular un comportamiento inteligente y, al igual que ella, está formado por unas entradas, que es por donde se reciben los valores, unidas a la salida a través de unos pesos que establecen su importancia. Con este mecanismo, se simula el comportamiento de las dendritas y la sinapsis. Para ajustar el comportamiento del perceptrón, en lugar de variar la cantidad de neurotransmisores, se varían los pesos de las conexiones, consiguiendo que la salida se active únicamente ante determinadas entradas (Aggarwal, 2023, pp. 5-6)[162]. Sin embargo, pese a sus innegables ventajas, las redes neuronales presentan un hándicap de cierto calado: requieren una elevada capacidad de cómputo, inasumible por los primeros ordenadores y que decantó la balanza hacia el uso de los modelos clásicos en las primeras investigaciones en este campo.

Desde los inicios de la IA, dos de los problemas que se han utilizado como campo de prueba para demostrar los avances han sido los juegos de tablero y el reconocimiento de patrones. En este sentido, el ajedrez ha sido uno de los temas más recurrente en las investigaciones de este campo[163]. Su interés reside en la gran cantidad de movimientos que se pueden hacer en

162 El perceptrón efectúa una función matemática que relaciona las entradas con la salida. Para ello, multiplica el valor de cada entrada por su peso asociado y suma todos los resultados. Si el valor obtenido es positivo, muestra un 1 en la salida, en caso contrario, muestra un -1.

163 Una máquina desarrollada por IBM llamada *Deep Blue*, venció en 1997 al campeón del mundo en aquel momento, Garri Kaspárov. Para una descripción detallada de su funcionamiento véase Campbell *et al.*, 2002.

Figura 2. Perceptrón



cada jugada como consecuencia del número de piezas que hay en el tablero y sus posibles posiciones. Los algoritmos inteligentes basan su funcionamiento en limitar esta explosión de combinaciones, ya que no es viable manejar tal volumen computacionalmente, y quedarse únicamente con la mejor jugada. Es lo que se denomina *búsqueda heurística*.

Así como en el caso del ajedrez se busca una forma inteligente de quedarse únicamente con los mejores movimientos, el reconocimiento de patrones consiste en clasificar imágenes, es decir, distinguir lo que aparece en ellas. Los ordenadores, desde hace años, superan a los humanos en los juegos de tablero por la facilidad de modelar las reglas que los rigen. El funcionamiento se basa en la lógica simbólica, que se fundamenta en el sistema aristotélico de silogismos. Parte de un conjunto de verdades absolutas que establecen el comportamiento y, mediante su procesado automático, se infiere nuevo conocimiento. Son modelos aptos para tareas abstractas y formales que ocurren en un universo cerrado y limitado, de manera similar al juego de ajedrez; sin embargo, la modelización de actividades aparentemente sencillas, como el reconocimiento de objetos, ha conllevado grandes dificultades por la cantidad de información intuitiva y subjetiva que requieren. *Logic Theorist* era capaz de razonar en base a un conjunto de axiomas que tenía preestablecidos, lo que se conoce como *base de conocimiento*. Esta aproximación obliga a formalizar el mundo que nos rodea y, dada su extensión y complejidad, limita el alcance de las tareas que el software es capaz de realizar, aunque contenga millones de reglas (Lenat, 1995, pp. 33-34). Dicha circunstancia hizo que se plantease la aproximación

opuesta, que la máquina fuese capaz de formar su propio conocimiento estableciendo patrones a partir de los datos en bruto, una capacidad que se denominó *aprendizaje automático* o *machine learning*.

> *Machine learning* is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and *learning* is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be *predictive* to make predictions in the future, or *descriptive* to gain knowledge from data, or both. (Alpaydin, 2014, p. 3)

A partir de este cambio de enfoque, se empezaron a abandonar los sistemas basados en reglas, dadas sus limitaciones, para enfocarse hacia la teoría de la estadística. Nacía la IA *subsimbólica* generalista que permite a los ordenadores afrontar problemas del mundo real y tomar decisiones en las que aparece la incertidumbre, tales como marcar el correo basura (Sahami *et al.*, 1998; Drucker *et al.*, 1999; Goodman *et al.*, 2007), diagnosticar el cáncer de mama (Osareh & Shadgar, 2010) o identificar caras de personas (Sharma *et al.*, 2020).

Las técnicas de *machine learning* más utilizadas en la práctica se basan en el *aprendizaje supervisado*. Se utilizan para problemas de clasificación y regresión[164] como los anteriormente citados. Para su funcionamiento, se le introduce al sistema un conjunto de muestras ya marcadas con la salida correcta para que modele su comportamiento automáticamente y minimice el error, es decir, vaya generando la función matemática que lo rige ajustando sus valores internos. Es un proceso que se va repitiendo sucesivamente hasta que se consigue una alta tasa de aciertos. Uno de los métodos más comunes que se emplean para ello es el *clasificador lineal*, denominado así porque separa el espacio de entrada en dos regiones con un hiperplano[165]. En el caso de trabajar en un espacio con dos variables, sería una recta. En un lado de la recta estarían las muestras que pertenecen a un conjunto, por ejemplo, el valor *SÍ*, y en el otro, las de otro conjunto, el *NO*. Con esta

164 La regresión es similar a la clasificación, pero la respuesta es continúa. Se utiliza para efectuar predicciones, como la edad de una persona según las películas que visiona o la evolución de los precios del mercado en base a las condiciones económicas de un momento dado.

165 Un hiperplano es un concepto geométrico que representa una forma de dividir un espacio en dos partes.

técnica, se consiguen sistemas que resultan de gran utilidad en la toma de decisiones, como ocurre con la clasificación del *spam*, concesión de créditos bancarios o inversión en bolsa. En el caso del correo basura, para entrenar al algoritmo se le suministra un conjunto de correos legítimos y no deseados ya clasificados, para que, automáticamente, el sistema ajuste sus valores internos a fin de conseguir la salida correcta. Si la muestra es suficientemente grande, será capaz de efectuar la clasificación con mínimos errores que, además, se pueden ir ajustando con el tiempo, dado que puede seguir aprendiendo. Es lo que ocurre cuando el usuario marca los correos que han ido erróneamente a la bandeja de *spam* como legítimos y a la inversa. En realidad, lo que está haciendo es ajustar el comportamiento del algoritmo para que, en un futuro, no vuelva a equivocarse.

En el caso de las técnicas que utilizan *aprendizaje no supervisado*, únicamente se requiere la introducción de un conjunto de datos representativo lo suficientemente grande para que, a partir de los ellos, el sistema forme una estructura sin intervención humana. A este proceso se le conoce como *knowledge discovery*[166]. Su hábitat natural es el campo del *big data* y la aplicación práctica por excelencia es el agrupamiento o *clustering*, que calcula la probabilidad de que un elemento pertenezca a un grupo (Murphy, 2012, pp. 3-18). Se ha utilizado con éxito para descubrir nuevos tipos de estrellas (Cheeseman *et al.*, 1996), crear publicidad dirigida según los hábitos de navegación (Berkhin, 2006) o comprimir imágenes (Sharma *et al.*, 2007). En el ámbito literario, se ha aplicado un algoritmo denominado LDA[167] en casos como el modelado de tópicos de prosa literaria decimonónica (Jockers & Mimno, 2013), del drama francés de la época clásica y de la ilustración (Schöch, 2017) y de la poesía del Siglo de Oro (Navarro Colorado, 2018).

El mayor problema de todas estas técnicas de aprendizaje automático con un enfoque probabilístico es que no se adaptan bien a las variaciones relevantes en la entrada, como pueden ser los cambios en la iluminación, en la posición de los objetos o en cuanto a los distintos acentos de las personas cuando se reconoce el habla. Todas ellas dependen fuertemente de los datos

166 Cuando se trabaja con una gran cantidad de datos, la definición de *data mining* es equivalente al *knowledge discovery*, aunque en ciencias de la computación es más habitual emplear este último.

167 *Latent Dirichlet Allocation* (LDA) es un modelo probabilístico generativo que se aplica a colecciones discretas de datos, como corpus textuales. "The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" (Blei *et al.*, 2003, p. 996), por lo que un documento será más probable que pertenezca a un grupo o tópico según las palabras que tenga.

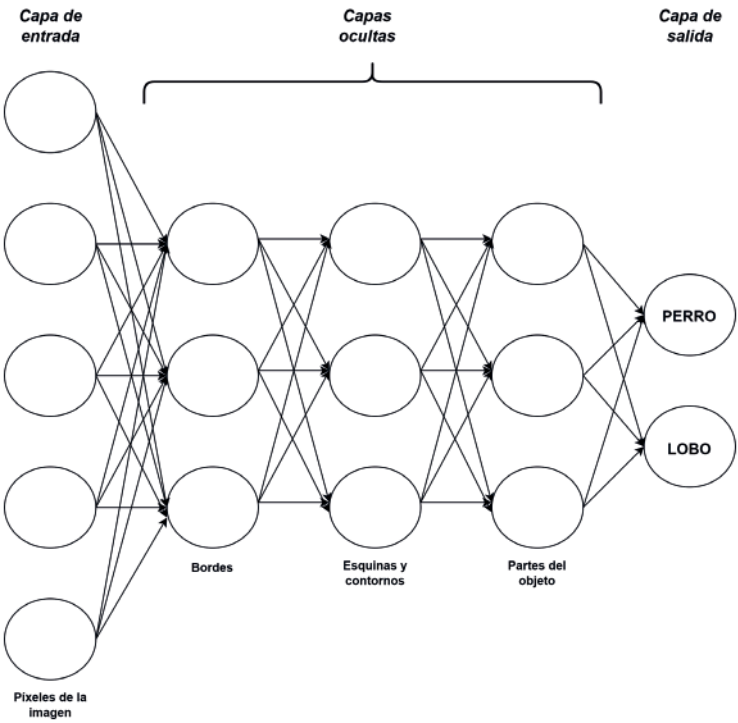
que se les introducen. En el caso de la clasificación del *spam*, se le indica qué campos del correo electrónico son relevantes para diferenciarlo, como el *asunto* o el *remitente*. A estos datos se les llama *características*, de manera que el sistema aprende cómo cada una de ellas se correlaciona con el resultado. Esto implica que a los sistemas basados en lógica subsimbólica, aunque construyen la función que los rige de manera autónoma a partir del entrenamiento con un conjunto de muestras ya clasificadas, se les debe indicar también las características que permiten discernir unas de otras. Esta fuerte dependencia condiciona la salida obtenida y provoca que tome especial relevancia su selección y representación, ya que, de lo contrario, el sistema no funcionará adecuadamente. En la práctica, un clasificador lineal etiquetaría dos imágenes con un perro en diferente posición en dos categorías distintas, mientras que, si tuviesen un perro y un lobo en la misma posición, las clasificaría como la misma.

La dificultad que conlleva modelar determinadas características manualmente, así como contemplar sus variaciones, impulsó la investigación de técnicas que permitiesen su detección automática. A esta nueva aproximación se la llamó *aprendizaje de la representación* y es en la que se basan las arquitecturas de aprendizaje profundo o *deep learning*, redes neuronales formadas por varias capas de células, cada una de ellas efectuando operaciones sencillas mediante cuya combinación el sistema modela funciones complejas de forma autónoma sin intervención humana. Asimismo, se consigue que el algoritmo sea insensible a variaciones irrelevantes, como el cambio en la posición de un elemento en una imagen, la variación de la iluminación o los distintos acentos de las personas (LeCun, 2015, pp. 437-438). La Figura 3 representa, de forma simplificada, un modelo de red neuronal artificial básico para identificar los objetos de una imagen, formado por dos capas visibles y tres ocultas. Las capas visibles son, por un lado, la entrada de los píxeles que forman la imagen y, por otro, el resultado. Las capas ocultas están en tres niveles con distintas funciones, cada una de ellas más compleja. En el primer nivel se detectan los bordes de los objetos de la imagen como primera aproximación para aislar los objetos. A continuación, se detectan las esquinas para formar los contornos de cada uno de los elementos y, finalmente, se unen las distintas partes para conformar e identificar cada una de las formas, ya en la última capa[168]. A esta estructura se le llama

168 Las distintas capas de la red neuronal también se pueden ver como instrucciones secuenciales de un algoritmo, por lo que, a más capas, podrá hacer operaciones más complejas. Una de las formas de medir la profundidad de un modelo, que toma como referencia esta interpretación, es en base al camino más largo a través del flujo de instrucciones que se ejecutan desde la entrada hasta la salida (Goodfellow, 2016, p. 8).

perceptrón multicapa (MLP) y tiene la virtud de crear, por sí sola, una función matemática compleja que asocia a un conjunto de valores de entrada otro de salida, a partir de la composición de funciones sencillas. El aprendizaje de la red, al igual que ocurre con el perceptrón, se consigue modificando los pesos de las conexiones entre las unidades de cómputo, simulando una sinapsis[169].

Figura 3. Modelo de red neuronal



169 "A multilayer network computes a nested composition of parameterized multi-variate functions. The overall function computed from the inputs to the outputs can be controlled very closely by the choice of parameters. The notion of learning refers to the setting of the parameters to make the overall function consistent with observed input-output pairs" (Aggarwal, 2023, p. 16).

Si bien el aprendizaje profundo se remonta, como hemos visto, a la propuesta de neurona artificial de McCulloch y Pitts en la que basa su funcionamiento, a lo largo del tiempo se le ha dado distintos nombres: en sus orígenes se le conoció como *cibernética*, en los años 80 se utilizó la palabra *conexionismo* para referirse al empleo de redes neuronales en la resolución de problemas y, a partir del 2006, tomó el nombre actual de *deep learning*. Sus modelos se inspiran en el funcionamiento de un cerebro biológico, tratando de imitar sus funciones. Dicho órgano es el responsable del comportamiento inteligente, por lo que, conceptualmente, si se desarrollan los principios computacionales que lo rigen y se duplica su funcionalidad, se conseguirá *inteligencia*. No obstante, en la actualidad, el término de *aprendizaje profundo* va más allá de esta perspectiva neurocientífica y se tiende a generalizarlo como un aprendizaje por medio de múltiples niveles de composición, de manera que, a partir de la combinación de operaciones sencillas, se crean otras nuevas más complejas (Goodfellow, 2016, pp. 12-14).

Una de las ventajas de las redes neuronales es que, a diferencia del clasificador lineal, no están limitadas a la división del espacio en hiperplanos. Esta característica permite que puedan modelar funciones más sofisticadas.

> The richness of the overall function computed by the neural network increases with the depth of the composition. Each layer of the neural network can learn successively more refined patterns from the patterns learned in previous layers. (Aggarwal, 2023, p. 122)

Cuanto más profunda sea la red, es decir, cuantas más capas tenga, aumentará la complejidad de la función que es capaz de calcular y su habilidad para crear intrincadas representaciones de datos. Esto conlleva a que sistemas como el de reconocimiento facial requieran de múltiples capas para conseguir resultados óptimos, puesto que una única capa tendrá dificultades para capturar las múltiples formas que componen una cara.

Aunque el incremento de la profundidad de una red neuronal se ve como una solución para aumentar la fiabilidad de la salida, no está exento de inconvenientes. El número de capas está directamente relacionado con la dificultad de entrenamiento, así como con la propia estabilidad de la red respecto a los parámetros de entrada. Obviando la demostración matemática de este comportamiento[170], lo que ocurre es que un pequeño cambio en la entrada —dado que afecta a una gran cantidad de capas— puede

170 Para la demostración matemática véase Aggarwal, 2023, pp. 122-129.

conllevar una modificación en las capas más profundas que desemboque en unas salidas erróneas. En la práctica, la elección del modelo final más adecuado se lleva a cabo mediante un mecanismo de prueba y error. Actualmente, existen librerías de software que facilitan la labor de crear redes neuronales de forma sencilla y que simplifican la tarea de simular distintas arquitecturas hasta obtener un sistema adecuado a los requisitos marcados.

Es indudable que la IA ha evolucionado desde los primeros años en los que demostraba teoremas matemáticos y jugaba de forma rudimentaria al ajedrez. Asistentes personales, diagnósticos médicos, recomendaciones de compra, concesión de créditos bancarios, inversión en bolsa y clasificación de *spam* son una pequeña representación de su aplicación práctica en el mundo real. Los encorsetados sistemas lógicos basados en reglas han dado paso al aprendizaje automático que, utilizando métodos probabilísticos, permite manejar la incertidumbre inherente en determinadas tareas, aparentemente sencillas para los humanos, pero tremendamente difíciles de modelar manualmente para que una máquina las ejecute.

El ejemplo práctico por excelencia de la IA aplicada al campo filológico es la traducción automática. Los sistemas actuales que han abrazado el aprendizaje profundo, como *Google Translate*, consiguen resultados cercanos a la perfección, aunque aún queda margen de mejora. Lo mismo ocurre con la clasificación de textos, generación de resúmenes y análisis de sentimientos, todos ellos pertenecientes al campo del procesamiento del lenguaje natural. En el campo de la visión por computador, se ha aumentado la precisión en la identificación de imágenes con un margen de error prácticamente despreciable. El reconocimiento de matrículas, la identificación de caras o la eliminación automática de objetos no deseados en una fotografía que acabamos de tomar con el móvil son algunos de los ejemplos de su aplicación en la vida cotidiana. El reconocimiento de grafías, en concreto, ha dado un salto cualitativo con el cambio del modelo estadístico clásico a las redes neuronales profundas. En un momento en el que la digitalización del patrimonio bibliográfico está plenamente instaurada, la siguiente fase lógica pasa por habilitar un mecanismo que extraiga su contenido textual con la mínima intervención manual. El reconocimiento automático de caracteres se presenta como el puente entre la digitalización y la representación textual del conocimiento.

4.2. Reconocimiento automático de caracteres

Los orígenes de las técnicas de reconocimiento automático de caracteres se remontan a principios del siglo XIX. El primer registro del que se tiene constancia es una patente de 1809 que describe un dispositivo dirigido a ayudar en la lectura a las personas con dificultades en la visión, aunque no recibió el impulso definitivo hasta seis décadas después, gracias al concepto del *escáner retina* de Charles R. Carey. Era un sistema de transmisión de imágenes que utilizaba fotocélulas para la captura de la información (Schantz, 1982, pp. 1-2). En 1900, Tuyrin se basó en esa idea para desarrollar un dispositivo también enfocado a facilitar la lectura a personas ciegas (Ning, 1993, p. 1).

El siguiente salto evolutivo en este campo se produjo ya en el siglo XX, a comienzos de los años 30, con las patentes de Emmanuel Goldberg (1932) y Paul H. Handel (1933) sobre unos dispositivos mecánicos que utilizaban las propiedades de la luz. Fue la primera vez en la que se aplicó la teoría de la probabilidad a este problema[171], un enfoque que aún se utiliza en las técnicas clásicas de OCR. Unos años después, entre 1938 y 1940, Vannevar Bush y su equipo construían, en el Massachusetts Institute of Technology (MIT), un selector de microfilms con un funcionamiento similar, que, desgraciadamente, se acabó desmantelado en plena II Guerra Mundial para aprovechar sus piezas en otras máquinas (Green, 1949, p. 350). La experiencia con este dispositivo y su relación directa con el diseño de ordenadores en los años que duró la contienda le sirvieron para proponer *Memex* (Bush, 1945, pp. 106-107), el dispositivo considerado por muchos como el precursor del funcionamiento de Internet.

Con la aparición de los ordenadores, se vio que el OCR tenía una utilidad directa en las labores empresariales: podía extraer información y clasificar documentos sin intervención manual. Esto, en realidad, venía a significar ahorro de personal y, por lo tanto, disminución de costes. IBM, el gran fabricante de máquinas empresariales en aquel momento, fue el primero que se adentró en este campo: en 1936, uno de sus investigadores publicaba la patente de una máquina de clasificación de cartas en base a su código postal[172]. Fue en los años 50, sin embargo, cuando, con la extensión de los computadores en las empresas e instituciones y

171 Ambos se basaban en la aplicación de la coincidencia del carácter a identificar con un patrón.

172 En la patente SU2056382A figuran como inventores del dispositivo Ayres Waldemar y Gilbert N. Fryer y la empresa a la que se le asigna es IBM.

el aumento de los datos que se trataban, se empezaron a poner en producción este tipo de dispositivos.

Pese al adelanto en la investigación que tenía IBM, la primera instalación comercial la hizo en 1954 otra empresa, IMR[173], en las oficinas de Nueva York de la revista *Reader's Digest*: un dispositivo que convertía documentos mecanuscritos en tarjetas perforadas para introducir la información directamente en el ordenador. Se consiguió reducir el trabajo de un mes a poco más de un día (Schantz, 1982, p. 10). Fue el espaldarazo definitivo a esta tecnología. Demostró que las máquinas podían leer y procesar documentos más rápido que una persona. A día de hoy, su uso se ha extendido a todos los campos: reconocimiento de matrículas, procesado de formularios, validación de pasaportes, conteo de billetes, conversión en tiempo real de escritura manuscrita y, por supuesto, transcripción de documentos y libros digitalizados, entre otros.

Descrito formalmente, un sistema de análisis de imágenes documentales, según Baird (2014, pp. 4-7), es aquel que se aplica a una imagen de una hoja de papel, formada por un conjunto de regiones textuales y sin texto, con el fin de extraer el mensaje que aloja. Las zonas de texto están compuestas por bloques o columnas que agrupan un conjunto de líneas horizontales o verticales según la lengua. El orden de lectura dentro de la línea también varía, pudiendo ser de izquierda a derecha o a la inversa, pero las letras siempre mantendrán el mismo orden que su transcripción vocal. Cada una de las líneas puede contener símbolos que representan las imágenes de las letras o palabras, así como los signos de puntuación. En los idiomas de Europa Occidental, convencionalmente se separan las palabras mediante un espacio en blanco, un elemento que, en ocasiones, dificulta su interpretación cuando su distancia se asemeja a la utilizada para marcar la separación entre las letras.

Esta descripción establece una relación directa entre símbolos lingüísticos y las correspondientes imágenes que los representan. Sin embargo, hay excepciones que desvirtúan esta regla, como es el caso de las ligaduras de caracteres o las contracciones, de uso habitual tanto en la escritura manuscrita como en los impresos incunables. Estos elementos aumentan el alfabeto base y dificultan la tarea del reconocimiento, dada su falta de estandarización, más allá de que puedan confundirse con errores caligráficos.

---

173 IMR es el nombre de la empresa que fundó David Shepard, un investigador del Departamento de Defensa de EEUU, después de construir una máquina capaz de reconocer las letras del alfabeto y que bautizó con el nombre de *Gismo* (Schantz, 1982, p. 8).

Actualmente, el análisis de imágenes documentales, comúnmente conocido como OCR, se encuadra dentro del área de reconocimiento de patrones, perteneciente al campo de la visión por computador, y acoge todos los temas relacionados con la identificación de formas y elementos dentro de imágenes. Es objeto de continua investigación desde los albores de la informática por la complejidad que entraña y la eterna necesidad de imitar los sentidos humanos para delegar tareas rutinarias y repetitivas en las máquinas.

El reconocimiento de caracteres está expuesto a un alto grado de subjetividad por las múltiples características y parámetros que definen un documento textual. En el caso de impresos, una misma grafía puede adquirir diversas formas según la tipografía utilizada[174], a lo que se une la posibilidad de variar el tamaño, el estilo y la separación entre letras y palabras, entre otras modificaciones. En los manuscritos, a todas estas dificultades se les une el propio modelo singular o único de la escritura de un copista, su idiosincrasia y las particularidades paleográficas de su *mano*, como se la suele denominar e, incluso, el diferente *ductus* de su letra según el momento de copia. Asimismo, la calidad de la imagen y/o del original del que se ha tomado también influirán en la correcta obtención del texto. De una parte, se encuentran los problemas provocados por un incorrecto proceso de captura, como poca resolución, mala iluminación, desenfoque o curvatura de la página; por otra, las propias particularidades materiales del original, ya sea en el soporte de escritura o en las grafías, como pueden ser manchas, borrones o agujeros, entre otros. El software deberá hacer frente a un cúmulo de problemas que tendrá que salvar para obtener un resultado que se acerque tanto como sea posible a la perfección o que, al menos, ayude a agilizar el trabajo humano, para que el tiempo invertido en su revisión sea menor que el requerido para efectuar directamente una transcripción manual. Aunque el objetivo a largo plazo sea la total automatización, mientras no sea posible, el éxito del sistema vendrá determinado por el ahorro de tiempo que le suponga al usuario su utilización y la facilidad de su manejo.

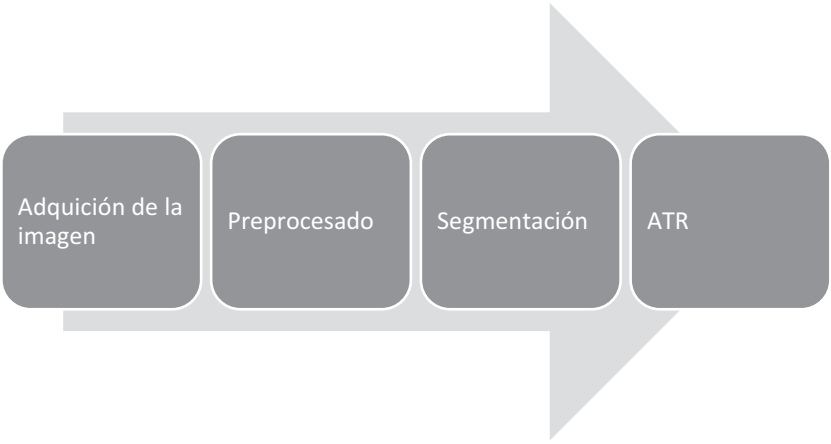
Lo más frecuente es que un OCR únicamente se centre en un modo de escritura, o bien impreso o bien manuscrito, dadas las diferentes características inherentes de uno y otro, sobre todo por lo que supone la diferente conectividad de caracteres, esto es, si es capaz de reconocerlos aislados o

174 Con el fin de facilitar la identificación de las grafías en materiales impresos por parte de los OCR, durante los 60 y 70 se estandarizaron dos tipografías: OCR-A y OCR-B (Baird & Tombre, 2014, p. 65).

unidos. Si el texto contiene ligaduras, habituales en la escritura manual, habrá que separar cada grafía previamente para identificarla según la técnica que se utilice. Esta característica inherente de los manuscritos conlleva la aplicación de complejas técnicas poco fiables tradicionalmente, por lo que, para evitarlas, surgieron los formularios con recuadros de división de letras que invitan a rellenarlos en mayúsculas, tan habituales en los procesos burocráticos, y que obligan a la separación de las letras.

El proceso general de OCR está compuesto por cuatro fases: adquisición de la imagen, preprocesado, segmentación y reconocimiento automático del texto (ATR). En la primera, se digitaliza el documento mediante una cámara fotográfica o un escáner con el fin de obtener una representación binaria del objeto físico para permitir su tratamiento computacional en las fases posteriores. Aunque es habitual pensar que, cuanta más resolución se utilice, se obtendrán mejores resultados, aumentarla demasiado también puede revelar la textura del papel, un elemento que, en el caso concreto del reconocimiento automático textual, no resulta productivo. Se debe alcanzar un compromiso entre resolución e información preservada[175].

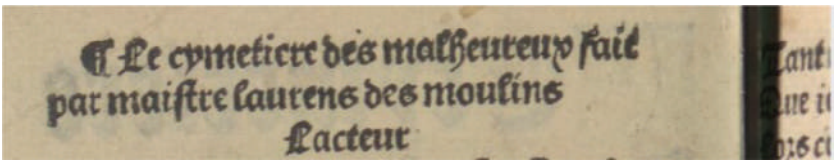
Figura 4. Proceso de OCR



175 Para calcular la resolución óptima, hay que tener en cuenta diversos factores, como el tamaño de la fuente del documento o si está en color o en blanco y negro. Algunos autores proponen que la línea más fina del documento debe tener al menos 3 píxeles de grosor una vez finalizado su tratamiento (Ha & Bunke, 1997, p. 5).

Las imágenes obtenidas del proceso de digitalización requieren de un tratamiento para lograr una captura lo más limpia posible y cercana a una situación ideal, lo que redundará en una mayor posibilidad de éxito. Ha y Bunke (1997, pp. 1-47) proponen efectuar un preprocesado en el que se somete la captura a una serie de transformaciones que dependerán del estado en el que se encuentre. Las primeras que se hacen son las geométricas, al ser habitual que la imagen tenga algún tipo de rotación o, en el caso de libros, deformación producida por la curvatura de la página[176]. La reconstrucción ideal debe ser una imagen perfecta sin distorsión a la que se le aplicará un filtrado con el objeto de acentuar la separación entre el fondo y las grafías. La complejidad, en este caso, vendrá determinada por la calidad de la captura y el estado del soporte. Una imagen con una alta resolución de un libro de reciente manufactura, con un fondo blanco uniforme y unos caracteres de imprenta perfectamente definidos en negro, no requerirá apenas tratamiento. Por el contrario, un incunable con el texto difuminado, el papel deteriorado con manchas, agujeros y dejando entrever el contenido del lado posterior de la página, necesitará la aplicación de varias técnicas consecutivas para paliar estos defectos con el fin de *eliminar el ruido*, como se conoce técnicamente (Shafait *et al.*, 2008a, p. 81). De esta manera, se conseguirá una localización más precisa de los elementos, entre los que figura uno de los más importantes, dado que permite encuadrar el texto: el marco de página. La mayor o menor complejidad de la tarea de establecer los márgenes y encuadrar el contenido vendrá determinada por la cantidad de elementos superpuestos a los ejemplares originales de los impresos, como sellos estampados de las bibliotecas o de procedencia de los fondos, anotaciones, dibujos o secciones de la página adyacente que no se ha eliminado correctamente en el proceso de digitalización, tal como se observa en la parte derecha de la Figura 5.

Figura 5. Laurent Desmoulins, *Le Cymetière des malheureux*, h. Aiv (Bibliothèque nacionale de France, RES-YE-1354)



176 Existen otro tipo de deformaciones, como las producidas por las lentes de captura, propias de digitalizaciones realizadas con equipos antiguos con ópticas deficientes.

Una vez finalizado el preprocesado, se efectúa la segmentación que "divide the document image into homogeneous zones, each consisting of only one physical layout structure (text, graphics, pictures, …)" (Shafait *et al.*, 2008b, p. 941). El objetivo último es diferenciar cada región de texto, marcas, anotaciones, imágenes y, en general, todos los elementos con entidad propia que conforman una página. Para ello, se utilizan dos enfoques: el ascendente y el descendente. La primera aproximación parte de la información local, los píxeles negros, para detectar las palabras y, a partir de ellas, formar las líneas y finalizar conformando los párrafos. Por otro lado, las técnicas descendentes utilizan la información global, como las bandas blancas de separación, para dividir la página en columnas, estas a su vez en párrafos, luego líneas y, finalmente, palabras (Lee & Ryu, 2001, p. 1241).

La localización de las palabras y sus correspondientes letras permite realizar el último paso de la transcripción: el reconocimiento automático de texto (ATR), una tarea que apareció antes que la informática y que, desde hace pocos años, tal como afirma Reul (2020, p. 36), ha cambiado radicalmente su enfoque. Tradicionalmente, se ha tomado como punto de partida la unidad mínima de la segmentación, la grafía, para extraer sus características y asignarla a una clase de carácter. Era habitual establecer la clasificación en base a determinados valores geométricos, tales como el tamaño en horizontal y vertical, el perímetro, el área y número de agujeros[177]. Esta forma de trabajo ha sido utilizada por todos los algoritmos de reconocimiento hasta versiones muy recientes, arrastrando un problema inherente: la necesidad de identificar cada letra individualmente. En realidad, si se partiese de imágenes con unas condiciones ideales del texto, con los caracteres totalmente definidos, con una separación perfecta, sin manchas ni borrones y un fondo homogéneo, no existiría ningún problema, pero las digitalizaciones distan mucho de esta situación, un problema que se acrecienta cuanto más antiguo es el documento original. No es extraño encontrar grafías incompletas en las primeras impresiones, o con contornos difuminados que tienden a unir las letras, dificultando su segmentación y, consecuentemente, su separación y correspondiente clasificación[178].

---

177 Entre las características geométricas más habituales figuran: tamaño horizontal y vertical y relación de aspecto entre ambos, perímetro, área, máxima y mínima distancia desde el borde hasta el centro de masas, número de agujeros, número de Euler, compactación y signaturas (Ha & Bunke, 1997, p. 41).

178 Pese a que no se utiliza la técnica de identificación de grafías individuales, Krichner *et al.* (2016) realizaron un trabajo de transcripción de incunables de un mismo impresor con un software que se basaba en ella. El volumen de trabajo que supuso lo hace inviable si se desea ampliar su rango de acción a diferentes tipografías.

La falta de precisión que tiene la segmentación para delimitar las formas de las grafías, tanto en impresos como en manuscritos, introduce una incertidumbre difícil de abordar si no es con el uso de modelos probabilísticos. Por tanto, se hace necesario recurrir a otro tipo de algoritmos que abandonen este enfoque y empleen el aprendizaje automático. Siguiendo esta corriente, inicialmente se recurrió a una técnica conocida como *modelo oculto de Márkov* o HMM, derivado de su nombre en inglés (Lu *et al.*, 1999) [179]. A grandes rasgos, su funcionamiento se basa en la toma de una u otra decisión en función de la probabilidad de su ocurrencia[180]. Sus dos campos de aplicación más habituales son el procesamiento del lenguaje natural, como en el caso, por ejemplo, del reconocimiento del habla (Jurafsky & Martin, 2009), y la bioinformática, donde se ha empleado con éxito, entre otras aplicaciones, para localizar genes dentro de la cadena del ADN (Schweikert *et al.*, 2009); sin embargo, tanto el primitivo sistema basado en segmentación como el enfoque con aprendizaje automático aplicando HMM son complejos de desarrollar y optimizar.

> Segmentation-based OCR systems require carefully designed character segmentation methods, since segmentation errors generally lead to errors in the output. In our experience, segmentation errors are the limiting factor for the performance of segmentation-based OCR systems. Segmentation-based OCR systems also require careful estimation of segmentation and classification costs; in particular, the fact that segments of different length compete for being present on the best path through the recognition lattice makes estimating costs difficult and may require heuristic tuning of the cost functions. Segmentation-free techniques like Hidden Markov Models (HMMs) applied to OCR avoid many of the difficulties of segmentation-based OCR systems, but still require careful choices of model structures, and face similar issues in heuristic modifications of their cost functions to achieve overall good performance. (Breuel *et al.*, 2013, p. 683).

Dado que el aprendizaje automático con técnicas estadísticas había demostrado ser una posible vía de mejora, a pesar de que aún tuviese detalles que dificultaban su optimización, el siguiente paso era recurrir al aprendizaje

---

179 *Hidden Markov Model*.

180 Para una descripción detallada y las bases matemáticas que rigen su funcionamiento véase Murphy, 2012, pp. 591-632.

profundo con redes neurales. Los primeros que lo llevaron a cabo fueron Graves *et al.* (2009, p. 865) con una propuesta en la que aplicaban una arquitectura específica llamada LSTM (Hochreiter & Schmidhuber, 1997)[181]. Consiguieron reconocer líneas de textos manuscritas con una reducción del error que alcanzó el 40% en algunos casos, respecto a las técnicas predecesoras que utilizaban HMM. En 2013, Breuel *et al.* modificaron esta idea y la aplicaron a líneas completas de textos impresos contemporáneos en inglés y a antiguos en alemán con letra *Fraktur*[182], una tipografía utilizada principalmente en Alemania con letras sistemáticamente separadas, en contraste con los tipos usados para una lengua románica. En el primer caso, se consiguió reducir el error a tan solo un 0.6% y en el segundo caso, el error fue del 0.82% (Breuel *et al.*, 2013, p. 685).

Pese a que se consiguió un gran avance utilizando LSTM, aún había posibilidad de mejora en determinadas áreas, como el margen de error, la velocidad de proceso y la simplificación de la fase de entrenamiento, que precisaba datos ya clasificados para empezar a funcionar. Esta evolución se produjo poco tiempo después de la mano del propio Breuel (2017, pp. 12-14), con su propuesta de incorporar otro tipo de redes neuronales, las CNN[183], que ya se habían utilizado con éxito en visión por computador (LeCunn *et al.*, 2010, p. 254), junto a las LSTM, combinándolas en una especie de modelo híbrido.

Actualmente, los motores de reconocimiento de texto de libre distribución, por los que no hay que pagar por su utilización, implementan esta arquitectura. Entre los que contemplan la transcripción de impresos antiguos, figuran *Calamari* (Wick *et al.*, 2020, p. 15) y *Kraken*[184], que, en este caso, además, permite combinar varios modelos de redes neuronales[185].

---

181 Para una descripción detallada del funcionamiento y las aplicaciones prácticas de la arquitectura LSTM véase Goodfellow *et al.*, 2015, pp. 397-400.

182 Según Reed (2019), *Fraktur* es un estilo de letra derivada de la tipografía utilizada por los primeros impresores. Pese a que a partir del xv se produjo una sucesiva romanización de tipos, en algunas partes del noreste de Europa siguió utilizándose el estilo gótico. Su rasgo diferenciador es la separación entre las grafías, aunque no está exenta de ligaduras. En Alemania se utilizó hasta bien entrado el siglo xx, hasta que fue prohibida por el gobierno Nazi en 1941.

183 Las CNN son un tipo de red neuronal profunda, con múltiples capas, que extrae las características que permiten la clasificación de forma automática. Para una descripción detallada de su funcionamiento, véase LeCun *et al.*, 1999.

184 https://kraken.re/main/index.html [consulta: 15/09/2023].

185 También existe *Tesseract*, financiado por Google, pero únicamente incorpora LSTM desde la versión 4.0, según su manual de usuario que se encuentra en la página del proyecto: https://tesseract-ocr.github.io/tessdoc/ [consulta: 15/09/2023].

Respecto a los OCR comerciales, aunque en el mercado existen varias empresas fabricantes, como ABBYY con su *FineReader*[186], que es el de uso más habitual en las bibliotecas, los que explícitamente facilitan su aplicación en textos antiguos y que han conseguido mayor difusión son *Transkribus*[187] y *Transkriptorium*[188], este último enfocado únicamente a manuscritos.

En definitiva, el proceso de OCR no se limita al reconocimiento de caracteres, sino que tan solo es un paso dentro de todo el flujo de trabajo que conlleva esta técnica. Para obtener resultados fiables, hay que partir de una imagen de calidad, con un buen preprocesado que nos mejore la nitidez de las grafías y su separación del fondo, algo esencial en el caso de material antiguo por su propensión a presentar deficiencias tanto en contenido como en el soporte. La aplicación de los tratamientos adecuados en este paso nos asegurará una segmentación precisa y, de esta manera, se podrán diferenciar claramente los márgenes, las distintas zonas del texto, las líneas que lo conforman y cada una de las grafías de las palabras. Todo ello para que el último paso, el reconocimiento de caracteres, obtenga el mínimo error posible. Al igual que ocurre con la clasificación automática entre imágenes de perros y lobos, las múltiples tipografías de los distintos impresos aumentan la complejidad de su identificación. Ha sido justo en este apartado donde la aplicación del aprendizaje profundo ha modificado totalmente la forma de abordarlo: se ha pasado de interpretar grafía a grafía, extrayendo las características de las formas, a reconocer palabras y frases enteras con la ayuda de redes neuronales que generan estas características automáticamente con el fin de identificarlas sin intervención humana. Para llegar a ello, previamente, el algoritmo deberá haber sido entrenado adecuadamente para que el modelo interno que construya se adecue a las normas de transcripción establecidas previamente. Como resultado de todo este proceso, a la salida y en condiciones ideales, se obtendrá el texto transcrito de la forma esperada. El siguiente paso natural será su interpretación, ya sea manual o computacionalmente, un campo de estudio que corresponde al *procesamiento de lenguaje natural*.

---

186 "While it is basically possible to train single glyphs and consequently a book-specific model using ABBYY, this is a tedious and ineffective task that seems to be mainly geared towards the recognition of quite specific ornament letters" (Reul *et al.*, 2019, p. 5).

187 https://readcoop.eu/transkribus/ [consulta: 15/09/2023].

188 http://www.transkriptorium.com/ [consulta: 15/09/2023].

4.3. La transcripción automática de textos medievales y del Siglo de Oro

La secuencia de pasos encadenados que da lugar a la transcripción automática del texto en formato digital se inicia con la captura de la imagen del objeto en cuestión, su transformación al mundo digital para que sus características puedan ser interpretadas computacionalmente, lo que suele llevarse a cabo de forma independiente con un software específico de tratamiento de imágenes[189]. Como resultado de la digitalización, se dispondrá de un conjunto de imágenes, bien separadas en varios archivos en algún formato gráfico como JPG, TIFF o PNG, o bien englobadas en un único documento, habitualmente un PDF. Dado que, a partir de ese momento, ya se dispone de la información en formato digital, los siguientes pasos se efectuarán íntegramente con el uso de un ordenador.

Para poder tener éxito en el reconocimiento del texto contenido en una imagen y, en especial, cuando se trabaja con material antiguo, es esencial llevar a cabo su preparación o preprocesado, un proceso durante el cual, básicamente, se aplicarán diversos filtros con el objetivo de eliminar manchas, transformar la imagen a blanco y negro, y dar nitidez a las grafías. El objetivo último de cada una de estas intervenciones es resaltar el texto respecto al soporte, generar un contraste y nitidez que facilite su identificación automática. De hecho, la correcta configuración de sus parámetros de funcionamiento será esencial para aumentar la tasa de éxito en las fases posteriores: la segmentación y el reconocimiento de caracteres. Aunque la segmentación implica, en principio, un solo paso, internamente está compuesta de varias operaciones, la primera de las cuales es la delimitación de las zonas de texto y su orden de lectura. En la Figura 6 se muestra un ejemplo de su resultado, en el que se observa que se han detectado las dos columnas y se les ha asignado un orden adecuado de lectura, primero la izquierda, que se marca con el número 1, y luego la derecha, marcada con el número el 2. Una vez aisladas las columnas, se lleva a cabo una nueva detección, pero esta vez de las líneas de texto que contienen y a las que, igualmente, se las numera secuencialmente según su lectura, con lo que se obtendrá, finalmente, el resultado mostrado en la Figura 7.

---

189 El software *ABBYY Finereader*, muy extendido entre las bibliotecas y que requiere licencia de uso, es de los pocos que integra el flujo completo de transcripción textual, incluido el primer paso de digitalización del documento, bajo la misma interfaz.

Figura 6. Detección de columnas (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)

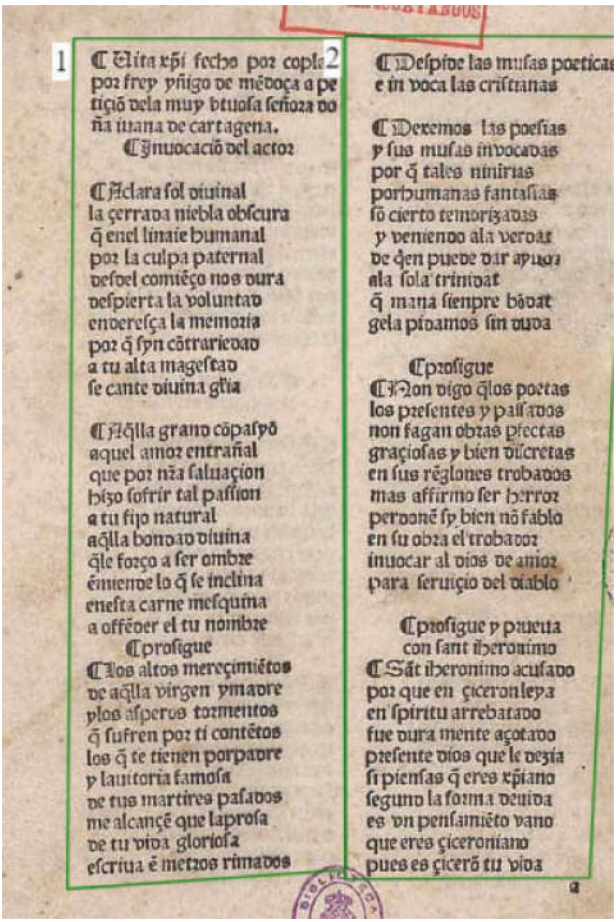
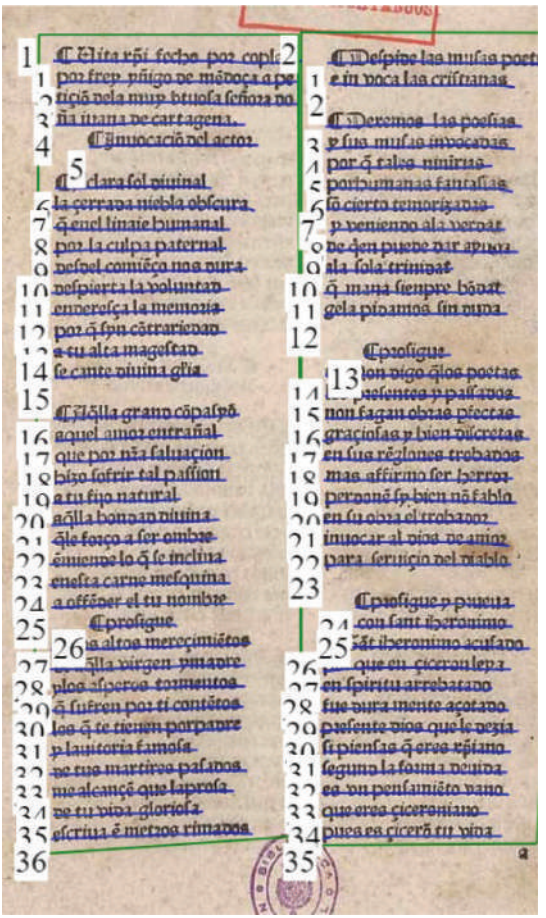
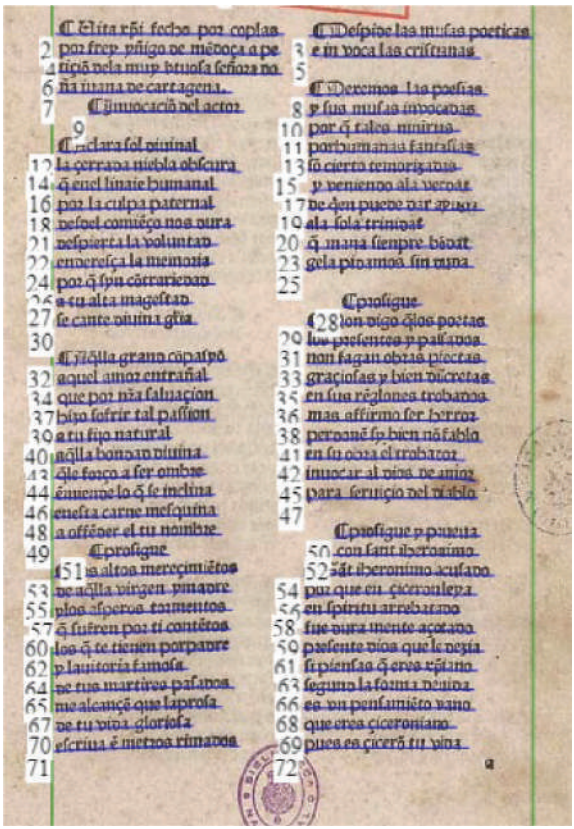


Figura 7. Detección de líneas (Biblioteca Nacional de España, INC/2159, h. aj[r] – 82IM)



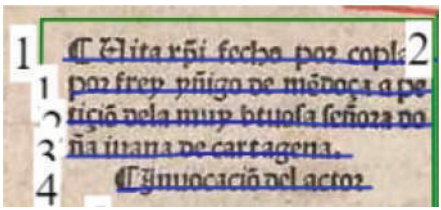
Si esta fase de segmentación presenta un funcionamiento incorrecto, el error se arrastrará hasta la fase final de transcripción, como se muestra en la Figura 8, en la que se detecta una única zona de texto, a línea tirada, ya que el software de segmentación ha confundido la separación entre las columnas con una cesura y, en consecuencia, el resultado de la transcripción de los versos no serán versos de arte menor, sino, aparentemente, versos de arte mayor formados por cesura y dos hemistiquios, con lo que ello implica en la secuencia textual.

Figura 8. Orden de líneas erróneo en h. aj^{r} (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)



El último paso, la transcripción propiamente dicha, se centrará en transformar las grafías en las correspondientes letras y palabras que representan, a partir de las líneas y columnas detectadas en el anterior proceso de segmentación, que aún son imágenes, tal como se muestra en el fragmento de la Figura 9, que ofrezco en paralelo al resultado de aplicarle un OCR adaptado a tipografía antigua. En este ejemplo concreto, se ha obtenido una edición paleográfica que ha mantenido las grafías originales, pero es posible entrenar el modelo para que obtenga una semipaleográfica que contemple la extensión de abreviaturas y la adaptación de grafías. Independientemente de la opción elegida, la salida será texto en bruto, sin una verificación sintáctica, léxica o semántica, por lo que es habitual que contenga errores menores. Este resultado, no obstante, se puede enriquecer con un postprocesado, que llevará a cabo una última verificación de errores con la aplicación de técnicas de procesamiento de lenguaje natural que corregirán, tomando como referencia un vocabulario y/o una gramática coetánea, el texto resultante.

Figura 9. Fragmento y transcripción (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM).



¶ Uita xp̃i fecho poꝛ coplas
poꝛ frey yñigo de mẽdoca a pe
tiçiõ dela muy btuoſa ſeñora do
ña iuana de cartagena.
¶ Inuocaciõ del actoꝛ

Aunque existe software específico para todos los procesos que se efectúan en este flujo de trabajo descrito, su utilización de forma individual conlleva un aumento de la complejidad de uso, dado que, al estar encadenados, se debe capturar de forma manual el resultado de cada paso para utilizarlo como entrada en el siguiente. Es por ello que han surgido herramientas que integran, bajo una misma interfaz, todos estos pasos unidos con la correspondiente aplicación software que lo gestiona, de manera que resulta transparente para el usuario la transición de una fase a otra. No obstante, pese a que en el mercado existen multitud de soluciones informáticas que hacen esta función, son pocas las que contemplan su utilización para la transcripción de impresos antiguos. La especialización y aún escasa

difusión de este tipo de transcripción conlleva que se aplique un sistema de reconocimiento que no está entrenado en grafías antiguas. Esta circunstancia se hace especialmente evidente en los errores de transcripción que se evidencian de la extensiva aplicación sistemática que están llevando a cabo las bibliotecas digitales utilizando un OCR adaptado a textos contemporáneos y sin una verificación del texto resultante, con errores léxicos derivados, en definitiva, del entrenamiento de los modelos que se utilizan en las redes neuronales del software de transcripción: "Let us mention this emblematic use case of *Gallica*: the word *budget* is often transcribed as *gadget* in 19th century press documents, long before that word even existed, which is clearly problematic" (Chiron *et al.*, 2017, p. 249). De esta manera, la tasa de reconocimiento del OCR empeora según la antigüedad de la edición, tal y como demuestran, a manera de ejemplo, estos resultados obtenidos a partir de cuatro impresos de siglos distintos escogidos aleatoriamente de *Gallica*:

Tabla 1. Tasa del OCR de *Gallica*

Ejemplar	**Fecha de edición**	**Tasa del OCR**
Le petit duc Charlotte Mary Yonge	1855	99.99%
Reflexions critiques sur les différentes écoles de peinture Jean-Baptiste de Boyer Argens	1752	98.55%
Socrate chrestien Jean-Louis de Balzac	1652	89.03%
La description de l'isle d'Utopie Thomas More	1550	87.85%

En efecto, dado que los modelos de inteligencia artificial han sido alimentados con textos contemporáneos, al aplicarlos sobre obras anteriores al siglo xx, se produce un sesgo en su interpretación como consecuencia de la variación diacrónica tanto del léxico como de la tipografía editorial que, además, intensifica su alienación acorde a la edad del impreso o manuscrito tratado.

> Historical typography alone poses a severe limit for the effectiveness of OCR: all available OCR engines have been extensively trained on modern fonts, but since historical fonts are very different from modern ones and the engines cannot be trained very well by

> end users on these historical fonts, training on modern fonts has very limited value for the OCR of historical printings. (Springmann & Lüdeling, 2017, p. 4)

Dicha limitación del software de transcripción se ha ido venciendo en los últimos años por el creciente interés entre la comunidad científica y la consecuente proliferación de modelos especialmente alimentados con documentación anterior al siglo xx, manuscrita o impresa. El material antiguo presenta una variedad de retos que no se encuentran en el contemporáneo; uno de ellos y el más evidente es su gran variedad tipográfica, pero no es el único. "As one moves earlier in the history of print, abbreviations become much more common and correspondingly more difficult to deal with" (Rydberg-Cox, 2009, p. 6). Asimismo, es habitual encontrarse con palabras que aparecen cortadas en dos líneas distintas con ausencia del guion de cambio de renglón. Pero no solo su contenido presenta características únicas, sino que el soporte que lo contiene también suele requerir un tratamiento diferenciado.

Si se obvia el primero paso del reconocimiento, esto es, la *captura* de la imagen en sí misma, dado que ya existen procedimientos operativos para digitalizar adecuadamente material antiguo[190], la siguiente fase, correspondiente al *preprocesado*, ya se enfrenta a diversas particularidades. Es habitual que el paso del tiempo y el tratamiento del que ha sido objeto el soporte haya dejado marcas evidentes de diversa índole. La ausencia de portada, los procesos de restauración para frenar el deterioro o la exposición a un ambiente húmedo, dependiendo de su severidad o de la destreza del restaurador, pueden perjudicar irremediablemente a la lectura del texto, tal como se aprecia en la Figura 10. En el caso de que se haya visto afectado por xilófagos, estos habrán dejado los característicos agujeros que delatan su presencia y que, en el peor de los casos, habrán afectado al contenido textual. Ello requerirá, a la hora de capturar las imágenes, tener la precaución de colocar una cartulina opaca detrás de la hoja afectada, con el fin de evitar la visibilidad de los caracteres de la hoja siguiente a través del agujero. En caso contrario, dicha circunstancia provocará que se generen errores en la transcripción, al verse las grafías de ambos textos entremezcladas, tal como ocurre en la Figura 11.

---

190 El Ministerio de Cultura y Deporte publicó en el 2021 el documento *Recomendaciones para proyectos de diglitalización de patrimonio bibliográfico y fotografía histórica* con indicaciones detalladas sobre todos los procesos implicados en la captura de material bibliográfico.

Figura 10. Zona deteriorada[191] (École nationale supérieure des Beaux-Arts, Masson 2055, f. IIII^{r} – 92VC)

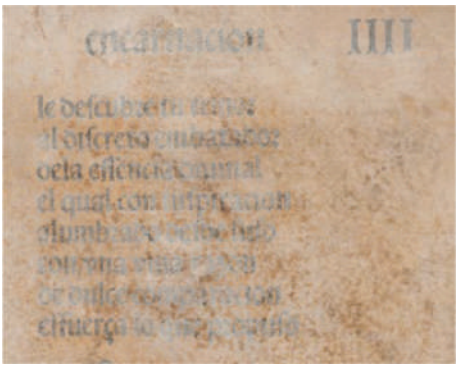
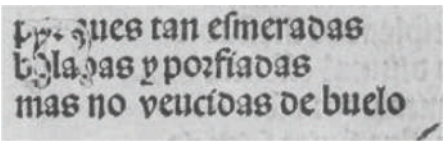


Figura 11. Agujero que afecta al contenido (Library of Congress, Incun. X.M52 PQ6180, f. V^{r} – 90*IM)



Los agentes ambientales no son los únicos que pueden afectar a la obra y su contenido. El nivel de eficiencia del OCR vendrá condicionado por la correcta localización de las zonas textuales que, en condiciones ideales, deberían presentar una estructura homogénea; sin embargo, el soporte puede haber sufrido alteraciones causadas explícitamente por el propietario, tales como dibujos, cancelaciones o notas manuscritas, como las que se observan en la Figura 12, que introducirán palabras ajenas al contenido original en el texto final. Algo similar ocurrirá si se han utilizado fragmentos de otros impresos diferentes en un proceso de restauración física del ejemplar, como

191 Parece una capa de celulosa aplicada sobre el texto en un mal proceso de restauración, aunque Marini considera que se trata solo de humedad y moho: "Presenta varias manchas de humedad y de moho, que no solo ha afectado el cartón debajo de la solapa en pergamino de la cubierta, sino también algunos folios del incunable, sobre todo los que han quedado del primer cuaderno. A causa de estas condiciones de conservación, en algunos puntos la tinta se encuentra muy descolorida, como es el caso del f. VIv, lo que a veces dificulta la inteligibilidad de los versos o de las imágenes" (Marini, 2023).

el que se muestra en la Figura 13. Al contener también texto, el software lo detectará como tal y, por tanto, segmentará el fragmento y lo transcribirá.

Figura 12. Anotaciones manuscritas (Biblioteca Nacional de España, INC/2900, f. V^{v} – 92VC)

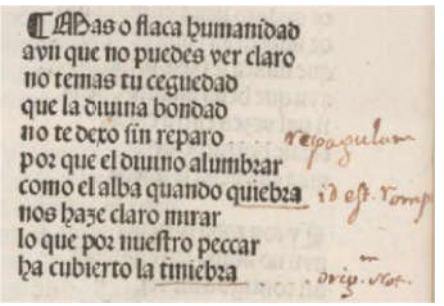
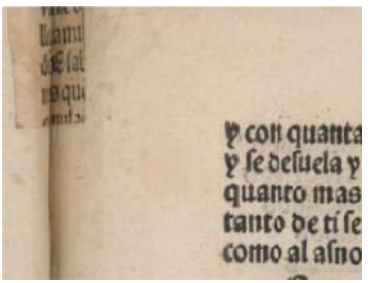


Figura 13. Restauración[192] con texto de otra obra (Biblioteca Nacional de España, INC/2900, h. aiiij^{v}-a5^{r} – 91*VC)



192 Para el refuerzo superior e inferior del cosido, en este ejemplar se utilizan escartivanas de papel de otro impreso en tipografía gótica, ya del siglo xvi, que Fernández Valladares ha identificado y datado, con un *terminus post quem*: "Por el diseño de algunas mayúsculas como M, D, E y algún calderón, pues ese diseño no se introducirá en la península hasta 1517 en la imprenta riojana de Arnao Guillén de Brocar (*cfr.* Norton, 1978, nº. 159, Tipo 10: 99 G). Aquí además se aprecia con una apertura mayor en el lóbulo derecho más redondeado, un diseño posterior al año 1520 del que dispusieron, por ejemplo, Miguel de Eguía en un cuerpo de 102 mm/20 lín., Juan de Brocar en dos fundiciones —una de 98 mm. y otra de 108— y Adrián de Amberes, sucesor de Miguel de Eguía en su imprenta de Estella (Casas del Álamo, 2016). Anoto la medida de 20 lín.: en la escartivana de la h. sign. Aj^{r}: 99-100 G; en la escartivana de la h. sign. Aiij^{r}-Aiiij^{v} = 98-99 G" (Fernández Valladares, 2019, p. 69, n. 47).

Todas estas dificultades confirman que, en realidad, cada uno de los pasos del flujo asociado a la transcripción se ve aún más condicionado cuando se trabaja con material antiguo. A los problemas inherentes del reconocimiento de grafías como consecuencia de su gran variedad y falta de homogeneidad, se unen las condiciones del soporte, cuya conservación habrá afectado a su contenido y, por ende, a la identificación de las grafías y a la consecuente extracción de su información textual.

Una de las primeras aproximaciones para vencer estos inconvenientes la llevó a cabo Ryndberg-Cox (2009, p. 13) con una propuesta de postcorrección manual del texto resultante, en concreto de las abreviaturas, asignando unos símbolos especiales que representaban los signos tironianos. Planteó un sistema mixto automático-manual que suplía las carencias en el reconocimiento de las formas pero que necesitaba una costosa revisión y corrección posterior por parte de una persona. Afortunadamente, con la aplicación del *deep learning* a la transcripción automática, que ya había producido un salto cualitativo en el reconocimiento de imágenes por parte de un computador, se consiguió reducir a un mínimo la intervención humana. Con las primeras técnicas que aparecieron, Dudczak *et al.* (2012, p. 94) entrenaron el OCR de libre distribución *Tesseract* a partir de muestras individuales de grafías con tipografía *Fraktur* sobre impresos comprendidos entre los siglos XVI y XVIII. Alcanzaron una exactitud de entre el 60% y el 80%, aunque aún precisaban de un trabajo previo en el que había que recortar manualmente cada letra del impreso, lo cual seguía siendo muy laborioso. Kirchner *et al.*, en un estudio centrado exclusivamente en incunables, consiguieron aumentar esta tasa con porcentajes cercanos al 95%, aunque la precisión de las palabras caía por debajo del 75% y el proceso de entrenamiento que utilizaron seguía basándose en una asociación de cada grafía con el símbolo correspondiente. Esta característica imposibilitaba generalizar el modelo para otros impresos antiguos dada la extensa variedad de tipos que existen. Sin embargo, pese a que utilizaron una técnica similar, es destacable la mejora en los resultados que consiguieron, fruto del tratamiento previo de las imágenes que hacían. Este hecho ponía en evidencia los problemas que tenían las digitalizaciones que se estaban realizando en las bibliotecas, con bajas resoluciones, alineaciones deficientes de líneas por capturas inclinadas o distorsiones por la curvatura de las páginas derivadas de la apertura del libro, entre muchas otras circunstancias.

> Obwohl inzwischen viele Inkunabeln und frühe Drucke digitalisiert sind und im Netz frei heruntergeladen werden können, reicht die hierbei zur Verfügung gestellte Qualität der Scans bzw. Bilddateien für ein hinreichend gutes OCRErgebnis oftmals nicht aus.

> Die Gründe hierfür reichen von schlechten Aufnahmebedingungen bis hin zu mangelnder Bildauflösung der frei zugänglichen Bilddateien.
>
> Ungeeignete Aufnahmebedingungen resultieren zum Beispiel in Verzerrungen im Buchfalz (vgl. Abb. 1) und nicht korrekt ausgerichteten Buchseiten. Eine weitere wichtige Voraussetzung für ein gutes OCR-Training ist die Auflösung der verwendeten Digitalisate. Bei im Netz frei verfügbaren Bildern liegt diese häufig unter 300 DPI, was die Erkennungsrate stark negativ beeinflusst. (Kirchner *et al.*, 2016, p. 179)

Por este motivo, destacaron la importancia de los dos pasos previos al reconocimiento de grafías si se partía de un escaneado deficiente: el preprocesado y la segmentación. Por primera vez, los datos demostraban la extrema importancia de utilizar una imagen limpia, con las grafías nítidas, el texto alineado correctamente y minimizando las distorsiones por la curvatura tipográfica de las imágenes.

La clásica asociación grafía-carácter que se venía empleando para generar los modelos de transcripción presentaba dos desventajas fundamentales: por un lado, limitaba la extensión de la aplicación del OCR a las tipografías utilizadas para su entrenamiento y, por otro, requería un arduo trabajo de recolección de cada una de las grafías. Springmann y Lüdeling cambiaron radicalmente este enfoque y propusieron partir de un corpus diacrónico con la transcripción paleográfica de impresos editados entre el siglo xv y principios del xx (Springmann & Lüdeling, 2017, pp. 13-18)[193]. En lugar de emplear letras individuales, entrenaron el modelo a partir de imágenes de líneas completas con su correspondiente transcripción. De esta manera, ya no solo se conseguía que la red neuronal generase una función que reconocía las grafías, sino también las palabras de manera íntegra. Para ello, tuvieron que cambiar la técnica de *deep learning* que se venía utilizando hasta ese momento y tomar como base redes neuronales recurrentes (RNN) de tipo LSTM. Con su propuesta, lograron una exactitud

193 Utilizaron el corpus RIDGES, que contiene tratados herbales escritos en alemán. Está formado por varias capas de anotaciones, cada una encuadrada en una categoría: léxica, sintáctica, morfológica, gráfica y, por último, la que contiene la transcripción. Se puede obtener más información en su página web https://www.laudatio-repository.org/browse/corpus/PySSCnMB7CArCQ9CNKFY/corpora [consulta: 25/11/2023].

que oscilaba entre el 76% y el 97% en la transcripción de palabras sobre un reducido corpus de impresos como objeto de estudio.

Esta novedosa aplicación de las RNN en la transcripción conllevó una nueva generación de OCR basados en *deep learning* que, ahora sí, obtenían resultados plenamente aprovechables sobre material antiguo. Una de las primeras aproximaciones se llevó a cabo dentro del *Proyecto Mambrino*, un grupo de investigación con sede en la Università degli Studi di Verona, que se dedica a los libros de caballerías y a las relaciones entre la Península Ibérica e Italia respecto de este género en prosa[194]. En su caso, crearon un primer modelo con la ayuda del software de OCR *Transkribus* utilizando un ejemplar de la *Segunda parte de Sferamundi*, decimotercer libro del ciclo de Amadís, impreso en letra cursiva por Michele Tramezzino en Venecia en 1560. Los resultados confirmaban el salto cualitativo que se había producido en la transcripción automática, puesto que alcanzaron una tasa de éxito cercana al 99% en el reconocimiento de caracteres aislados (Bazzaco, 2018, pp. 268-269). Ante el resultado obtenido con este software, la comunidad científica se animó a embarcarse en proyectos más ambiciosos que venciesen su principal limitación: su utilización de forma extensiva en otros impresos. Este modelo, que se denomina *individual* por haberse entrenado con una única edición, obtendrá su máximo rendimiento con la tipografía de ese impreso en concreto. Es por ello que se decidió crear un modelo que abarcase un conjunto de ediciones y obras con el objetivo de que fuese aplicable a otros impresos coetáneos con diferente tipografía, esto es, un modelo *extendido*.

Dada la carencia de modelos de transcripción adaptados a las tipografías de los impresos españoles de la Edad Moderna, surgieron dos nuevas propuestas, una para letra gótica y otra para redonda, centradas exclusivamente en *Transkribus*, ya que una prueba de concepto basada únicamente en 1500 palabras había obtenido una tasa de error por carácter (CER) cercana al 2% en ambos casos. Estos modelos obtuvieron de nuevo una fiabilidad en torno al 99% (Bazzaco *et al.*, 2022, pp. 92-95) y se encuentran actualmente disponibles para utilizar públicamente dentro de *Transkribus*

194 "El grupo nació en 2003 con la ambición de estudiar el *corpus* de libros de caballerías italianos producidos en Venecia a mediados del siglo XVI (1546-1568) como traducción e imitación del género caballeresco hispánico que, a partir del éxito del *Amadís de Gaula* de Garci Rodríguez de Montalvo (1508), se desarrolló en una avalancha de imitaciones e innovaciones, con ciclos completos de novelas en constante mutación, y crearon el hábito de leer ficción, involucrando varias generaciones de lectores apasionados: los hermanos de Don Quijote" (Neri, 2019, p. 444).

con el nombre de *Spanish Gothic* y *Spanish Redonda* respectivamente. El *dataset*, formado por dieciséis textos, abarca libros impresos desde el 1487 hasta el 1527 en letra gótica, mayormente de caballerías[195]. Para el modelo en letra redonda, se han empleado quince textos histórico-caballerescos impresos entre 1578 y 1607 y una decena de relaciones de sucesos que abarcan desde el 1598 al 1663[196].

Con el fin de comparar la mejora de rendimiento que se consigue con la especialización de los modelos individuales respecto a los extendidos, se generó uno nuevo entrenado, en esta ocasión, a partir de un impreso en gótica, en prosa y con puesta en página a doble columna: la edición de *Silves de la Selva* publicada en 1549 en el taller de Dominico de Robertis situado en Sevilla. Igualmente, se utilizó *Transkribus* como software de base, empleando las primeras veinte páginas del impreso para el entrenamiento y con unos criterios de edición enfocados a obtener una edición diplomática. El modelo resultante, ya entrenado, transcribió el resto de la obra con unos catorce errores por página. No obstante, pese a que su especialización podría inducir a pensar que sus resultados serían mejores que con uno extendido entrenado con otras obras coetáneas, la realidad fue muy distinta. Al aplicarle el *Spanish Gothic* anteriormente comentado, consiguieron una media de siete errores por página, logrando una exactitud del 99.78% y demostrando, por tanto, la supremacía de los modelos más generales (Blasut, 2022, p. 185).

Todas estas investigaciones se habían llevado a cabo con *Transkribus*, un software de transcripción inicialmente llamado *Transcriptorium*, fruto de un proyecto con financiación de fondos europeos[197] para crear un sistema de uso abierto y cooperativo que, poco después, pasó a ser un producto

195 Se incluyen tres obras del periodo incunable, *Doctrinal de los Caballeros*, impresa en Burgos en 1487 por Fadrique de Basilea, *La Fiameta* de 1497, impresa en Salamanca por el mismo taller que la *Gramática de Nebrija* y *Crónica del Rey Don Rodrigo*, impresa en Sevila en 1499 por Meinardo Ungut y Estanislao Polono.

196 Para los criterios de transcripción, véase Bazzaco *et al.*, 2022, p. 95.

197 "The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600707 – tranScriptorium" (Sánchez *et al.*, 2013, p. 2). "This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No 600707 and Horizon 2020 research and innovation programme under grant agreement No 674943" (Kahle, 2017, p. 23).

comercial[198]. Dado este cambio radical en su política de uso, que impedía su acceso de forma gratuita, surgieron otras alternativas de libre distribución, como es el caso de *eScriptorium*, basado en el motor de reconocimiento *Kraken* (Kiessling *et al.*, 2019, pp. 21-22). Al igual que *Transkribus*, consta de una parte para la segmentación con el objetivo de hacerla automática y otra que contempla la transcripción. En ambos casos, existe la posibilidad de entrenar un modelo adecuado al material que se vaya a procesar, aunque también tiene la posibilidad de realizar ambos pasos de forma manual o semiautomática. "It is highly modular, and each module has a large number of parameters that the user can set to accommodate the specific needs of the case in question" (Stokes *et al.*, 2021, p. 3) y, dado que el código fuente está disponible al ser de libre distribución, el usuario lo puede adaptar a sus requisitos si tiene conocimientos de programación. *Kraken* también ha sido utilizado por *OCR4all*, otra de las plataformas que buscan integrar todos los procesos del flujo del OCR. En particular, lo utiliza exclusivamente para la parte de segmentación, dado que para la parte de transcripción emplea otro software de libre distribución: *Calamari*.

Este incremento de alternativas de software de transcripción ha favorecido la aparición de los primeros estudios comparativos. Tanto *Transkribus* como *OCR4all* se han aplicado en paralelo a un conjunto de impresos de Arnao Guillén de Brocar, lo que le confiere una cierta homogeneidad como corpus tipográfico[199]. Para su entrenamiento, se seleccionaron ediciones de principios del XVI, en su mayor parte latinas[200], con tipografía gótica y redonda, a partir de los cuales se generaron diversos modelos variando parámetros y textos base. Los resultados no lograron decantar la balanza hacia ninguno de los dos softwares, ya que, aunque *Transkribus* obtenía mejores resultados en la transcripción semidiplomática, *OCR4all* era superior en la diplomática pese a que el número de líneas utilizadas para generar su modelo había sido mucho menor que el utilizado para *Transkribus*. Dada la

198 "Transkribus originated from an EU FP7 funded project 'Transcriptorium', and then from an EU Horizon 2020 funded project, READ (Recognition and Enrichment of Archival Documents), which launched an online HTR tool in 2015. It has since been developed further by the READ-COOP, structured around a cooperative of partner institutions and becoming a paid-for service in 2020" (Nockels *et al.*, 2022, p. 368).

199 "Pese a tratarse de impresos con características tipobibliográficas muy diversas, tienen el nexo común de haberse concebido en el taller dirigido por el mismo maestro impresor con uso de letrerías, disposición de página y convenciones editoriales repetidas" (Ayuso García, 2022, p. 154).

200 Para obtener un detalle de las obras seleccionadas, véase Ayuso García, 2022, pp. 165-166.

indefinición obtenida, estos hechos no permiten elegir entre uno y otro de forma objetiva, por lo que precisan de un mayor estudio y "deben mejorarse usando las herramientas que ambos sistemas proporcionan para poder ser útiles en filología" (Ayuso García, 2022, p. 164).

Los corpus son la base del entrenamiento para generar nuevos modelos y, ante la falta de disponibilidad o restricciones en su uso que presentaban las investigaciones previas sobre español medieval, Gille Levenson compiló un *dataset* con la ayuda de la aplicación *eScriptorium* y lo publicó bajo licencia CC-BY-NC-SA. Su contenido está extraído de la traducción castellana del *Regimiento de principes* de Egidio Romano —en concreto, del primero y el tercero de los tres volúmenes que componen la obra—. El corpus está compuesto por el texto de 318 folios del incunable sevillano salido del taller de Ungut y Polono en 1494 y el de diez manuscritos que se conservan de la misma obra[201], todo ello con la idea de generar un nuevo modelo extendido, con el que consiguió una exactitud del 96.30% en su aplicación a manuscritos fechados entre el siglo xiii y el xv.

> The models are usable to pre-annotate documents, in order to produce new data faster. As for scholarly editing, they probably need some prior finetuning to be perfectly adapted to a particular hand. Moreover, but they should be usable, for distant reading: the current accuracy is good enough to produce good results in stylometry, for instance. (Gille Levenson, 2023, p. 10)

Este enfoque demuestra que, pese a la aparente limitación léxica que tiene el modelo al haber sido entrenado con una única obra, la variedad tipográfica consecuente de sus distintos testimonios permite extender su aplicación con resultados aprovechables para determinadas tareas y, evidentemente, con un entrenamiento más profundo, es posible aumentar su grado de precisión, hasta el punto de que se puede tomar como base para una edición, sea diplomática o crítica.

Uno de los últimos modelos castellanos aparecidos en *Transkribus* adaptado a textos antiguos es el denominado *Coloso Español*[202]. Es un modelo extendido entrenado con más de 11 millones de palabras y que, según

---

201 Las páginas concretas utilizadas de los manuscritos y del incunable se pueden consultar en Gille Levenson, 2023, p. 4.

202 Se pueden consultar sus especificaciones en la página web de *Transkribus*, en la dirección https://readcoop.eu/model/coloso-espanol/ [consulta: 10/11/2023].

su descripción, está diseñado para transcribir una gran variedad de textos, desde manuscritos medievales hasta documentos del siglo xx. No obstante, pese a esta gran magnitud de datos, las primeras pruebas realizadas por investigadores no han evidenciado un comportamiento acorde a lo esperable por su envergadura, de lo que deja constancia Fradejas Rueda, quien ha "probado este Coloso español con varios folios de varios manuscritos castellanos y los resultados han sido muy pobres, cuando no directamente desastrosos" (2023, p. 16).

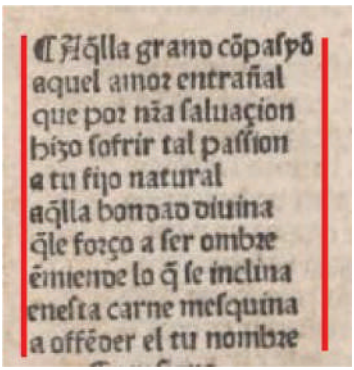
De todo ello se concluye, por una parte, que los modelos individuales únicamente entrenados con un impreso y, por ende, con una tipografía concreta, presentan escasas o nulas ventajas, mientras que, por otra, los primeros resultados obtenidos con los modelos extendidos sobreentrenados tampoco evidencian un aumento significativo de rendimiento, sino todo lo contrario. Por tanto, todo apunta a que el camino a seguir son los modelos extendidos especializados en periodos temporales concretos. No obstante, es necesario distinguir que, dentro de estos, existen dos enfoques: bien utilizar impresos de distintas obras de diversos impresores, o bien partir de una variedad de testimonios de la misma obra. El objetivo de ambas perspectivas es, al fin y al cabo, el mismo: lograr una variedad tipográfica que permita su carácter generalista. Las investigaciones avalan que ambas opciones consiguen buenos resultados y generan modelos de transcripción aplicables más allá de los textos con los que han sido entrenados.

5. La transcripción automática de incunables poéticos

Las diversas iniciativas de creación de modelos informáticos capaces de procesar material antiguo están propiciando que su transcripción automática sea cada vez más precisa; sin embargo, el gran volumen de impresores y su diversidad tipográfica, así como las diferentes puestas en página de las ediciones resultantes, siguen dificultando su aplicación a gran escala sin la intervención humana. Más allá del estado en el que se encuentre el soporte, son estos los principales problemas a los que nos enfrentamos en el reconocimiento automático de impresos, tanto en lo que respecta a la correcta interpretación de las grafías góticas, como en relación a los propios procesos de segmentación, es decir, a la división de las zonas textuales en filas y columnas, aspecto este último que se complica en los incunables poéticos, por la distribución estrófica.

En efecto, por un lado, las tipografías incunables han sido talladas manualmente por diferentes artesanos e imitando, con diferentes diseños, la escritura manuscrita, por lo que, prácticamente, acaban presentando la misma complejidad y variedad que esta, de manera que su interpretación y reconocimiento morfológico es uno de los hándicaps del proceso de transcripción automática de la poesía impresa incunable. Por otro lado, la puesta en página de este tipo de material, pese a que pueda mantener la coherencia a lo largo de la edición, es compleja por su habitual distribución en columnas con márgenes no paralelos, tal como se muestra en la Figura 14. Aunque es habitual que el margen izquierdo se mantenga alineado, en el caso de los incunables poéticos no ocurre lo mismo con el derecho, que, como es lógico, queda descuadrado, pues unos mismos pies métricos no implican un idéntico número de grafías, que, por otro lado, tampoco responden a tipos de unas mismas dimensiones. Esta singular característica de la poesía dificultará la detección correcta de las zonas textuales por parte de los algoritmos y, por tanto, incrementará las posibilidades de obtener una segmentación errónea.

Figura 14. Márgenes de una estrofa (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)



A pesar de que estas son, en efecto, las principales cuestiones a abordar, de manera general, es cierto que no son las únicas y debemos considerar otras circunstancias, como es el caso de la existencia de rúbricas, habitualmente sangradas y, en ocasiones, incluso, con distinta tipografía, con el consecuente aumento de confusión en el algoritmo de detección; o de las xilografías insertadas entre el texto; o de los titulillos en la cabecera de página y, normalmente, centrados, sin alinear con el resto del contenido. A toda esta lista de contratiempos y dificultades, habría que añadir, además, las habituales problemáticas causadas por el estado del soporte ya enumeradas en apartados anteriores como agujeros, restauraciones, manchas, borrones, notas al margen y sellos de poseedores, entre otras.

La inteligencia artificial ya se ha aplicado con éxito en circunstancias similares, por lo que parece el camino a seguir también en este caso. No obstante, la creación de un sistema que sea capaz de transcribir un incunable poético con la exactitud suficiente para aprovechar su resultado en la elaboración de una edición crítica o paleográfica[203] aún es un campo inexplorado y, en consecuencia, es la razón que justifica el principal objetivo de esta monografía: crear un modelo extendido para una inteligencia artificial basado en incunables poéticos, atendiendo a sus principales particularidades técnicas, a fin de que facilite la labor filológica de edición de cancioneros.

---

203 Si no otras, puesto que se le pueden ofrecer criterios modernizadores, por ejemplo, para obtener, incluso, una edición actualizada, de alta divulgación, aunque, aun en este caso, también deberá tener la revisión crítica del filólogo, en última instancia.

5.1. Delimitación del corpus

La primera cuestión a abordar en la construcción de un sistema que reconozca grafías de forma automática es la selección del corpus con el que será entrenado. La creación de un modelo individual, es decir, a partir del entrenamiento de una red neuronal con un único documento, implicará que únicamente será capaz de reconocer eficazmente textos que compartan esa misma tipografía. Si se trata de textos contemporáneos, la problemática es prácticamente inexistente al existir modelos extensamente entrenados con las tipografías de imprenta actuales que, asimismo, no varían sus grafías dentro de la misma familia gracias al uso de la informática. Si maquetamos un documento con *Arial*, sabemos con exactitud su apariencia final y no cambiará independientemente del software o dispositivo que utilicemos para representarla. Sin embargo, su aplicación en material antiguo dista de ser óptima por el motivo contrario: la gran diversidad y falta de homogeneidad tipográfica existente derivada de su construcción artesanal. Este razonamiento inclina la balanza, sin lugar a dudas, hacia un interés por los modelos extendidos, aquellos entrenados con tal variedad de tipos que permiten extrapolar su uso más allá de los impresos con los que han sido entrenados. No obstante, y tal como se ha comprobado con los enormes modelos extendidos que han surgido recientemente, alimentados con millones de palabras, como es el caso del *Coloso Español*, más cantidad no significa necesariamente más calidad (Fradejas Rueda, 2023, p. 16). Hay que encontrar un punto óptimo en el volumen de datos suministrados al modelo para no sobreentrenar la red neuronal más allá de los límites para los que ha sido diseñada; en caso contrario, se puede provocar un comportamiento errático y totalmente opuesto a los objetivos buscados.

Para el entrenamiento de los modelos extendidos se puede optar por dos enfoques: basarse en una única obra con ediciones de diversos talleres o bien utilizar un conjunto de obras distintas con diversa tipografía. Independientemente del método elegido, el objetivo que se persigue es el mismo: conseguir un abanico de tipos representativo de la época en cuestión, que, en nuestro caso, sería el siglo xv. Las investigaciones llevadas a cabo hasta la fecha (Bazzaco *et al.*, 2022; Gille Levenson, 2023) han demostrado que tanto una opción como la otra son perfectamente válidas para su aplicación en material antiguo y deberían permitir entrenar una red neuronal para que generase una transcripción aprovechable como punto de partida para una edición crítica. Estas conclusiones vienen a demostrar que, más allá de la variedad léxica, el corpus utilizado debe abarcar una amplia selección tipográfica que permita a la red neuronal modelar correctamente una función que asocie la misma grafía a las diversas letras de molde que la representan.

El corpus para el entrenamiento y establecimiento del modelo, evidentemente, debe ser contemporáneo a los documentos que se pretendan transcribir. Nuestro objeto de estudio, por el marco científico en el que se encuadra esta monografía[204], son los impresos poéticos incunables del contexto hispánico, que se inauguran con la edición temprana de *Les trobes en lahors de la Verge Maria*, clasificada como 74*LV por Dutton, que se imprimió en 1474[205]. Este primer cancionero impreso se distingue por su poliglotismo[206], puesto que, además de ser de autores varios, recoge, junto a las poesías en catalán, una en toscano y cuatro en castellano[207].

No obstante, pese a su temprana impresión —y, en parte, precisamente por ello—, *Les trobes* son un caso aislado[208] dentro de la poesía castellana de cancionero, representada aquí por solo cuatro composiciones y con una tipografía redonda o romana, que contrasta con la eclosión de la gótica poco después, que caracterizará el resto de incunables poéticos. Es esta la razón fundamental por la que dejamos de lado esta edición como corpus para entrenar una red neuronal, puesto que, por su variedad y alcance, focalizaremos los incunables poéticos impresos en tipografía gótica[209].

---

204 Tanto el proyecto de investigación en que se enmarca (*Poesía, Ecdótica e Imprenta*) y el grupo de investigación que lo desarrolla (CIM – Cancioneros Impresos y Manuscritos, https://www.cancioneros.org [consulta: 10/12/2023]).

205 A pesar de que el colofón "no lo conserva el ejemplar único, pero hay evidencias de que lo llegó a tener en su sección mútila final, porque es evidente que lo vio Josep Rodríguez, pues, si no hubiese sido así, no habría indicado que se imprime 'En Valencia 1474', pero que 'Falta nombre de Impressor'. ¿De dónde falta? ¿Por qué iba a aportar este matiz de ausencia, frente a la constatación de los otros dos datos, si no tuviese delante un colofón? Este colofón que describe Josep Rodríguez, con lugar y año, corresponde exactamente al modelo que encontramos en los dos primeros incunables valencianos que lo contenían, ambos con la misma tipografía que *Les trobes en lahors de la Verge Maria* y atribuidos tradicionalmente a Lambert Palmart" (Martos, 2022b). Véase también Martos, 2023a, pp. 79-87.

206 Para un estudio detallado de este incunable, véase Martos, 2023a.

207 "Y no tres, como computaron erróneamente José Mª Torres Belda (1874, p. 47) e, incluso, el propio Martí de Riquer, que olvidó el poema de Pere de Civillar en su catalogación de las poesías en castellano" (Martos, 2023b, p. 137).

208 Excepción hecha de la *Obra de la sacratíssima Conceptió de la intemerada mare de Déu*, impresa en Valencia por Lambert Palmert el 1487 y que contiene un poema en castellano de Juan Tallante, cuya primera edición crítica ha ofrecido recientemente Josep Lluís Martos (2024b).

209 Aunque no solo ella, si bien es la principal, porque el corpus de poesía castellana también es muy limitado.

La verdadera difusión impresa de la poesía castellana de cancionero, a una cierta escala, comenzaría, por tanto, en el taller zamorano de Antón de Centenera, con la impresión del pliego suelto del *Regimiento de príncipes* de Gómez Manrique (82*GM) y con la *editio princeps* de las *Coplas de la vita Christi* de fray Íñigo de Mendoza, que se remata con una obra más breve, como es el *Sermón trobado* de este mismo autor (82IM)[210]. Esta última edición marcó un punto de inflexión en la manera de distribuirse y consumirse la poesía[211]:

> Se ha relacionado la aparición del primer pliego suelto poético que conocemos, el *Regimiento de príncipes* de Gómez Manrique (82*GM), con el éxito editorial de la *editio princeps* de las *Coplas de la vita Christi* de Íñigo de Mendoza (82IM). Esta primera edición inauguraba ese modelo editorial de obras poéticas al que me refería anteriormente, que acompañaba una principal de otra complementaria, de menor extensión —en este caso el *Sermón trovado*, también de fray Íñigo—, que funcionaba a menudo como remate cuantitativo y cualitativo del producto comercial. (Martos, 2018b, p. 527)

Las *Coplas de la vita Christi*, sin duda por los orígenes nobles de fray Íñigo y por su relación con la corte de la reina Católica, fue una de las obras poéticas más leídas en la segunda mitad del xv, a la luz de su transmisión manuscrita y, en lo que respecta a los objetivos de esta investigación, un verdadero *best-seller*, puesto que se conocen hasta ocho incunables.

El testimonio más antiguo conservado de las *Coplas de la vita Christi* de que se tiene constancia es el *Cancionero de Oñate-Castañeda*, catalogado como HH1 por Dutton y conservado actualmente en la *Houghton Library* de Harvard University[212]. En este cancionero manuscrito se recopila un total

---

210 Cuya transmisión y edición crítica, a la luz de todos los testimonios, ha ofrecido recientemente Francisco Crosas (2024).

211 "Durante la época incunable, la poesía se vehiculó en soportes breves y, por tanto, de difícil conservación" (Martos, 2018b, p. 527).

212 Para la descripción codicológica del *Cancionero de Oñate-Castañeda*, y su transcripción completa, véase Severin, 1990. La versión digitalizada del manuscrito se puede consultar en línea en la Biblioteca Virtual Miguel de Cervantes a través de la dirección https://www.cervantesvirtual.com/obra-visor/cancionero-de-onate-castaneda-hh1/html/ [consulta: 10/09/2023].

de 92 poemas[213], según la clasificación de Garcia (1990, pp. IX-XV), agrupados por temática: "La *Vita Christi* de Fray Íñigo de Mendoza (56) inaugura otra serie inspirada por la Pasión, completada con la *Pasión trobada* de Diego de San Pedro y las *Coplas a la Varonica* de Fray Ambrosio Montesino" (1990, p. XX) [214].

Es un poema que ha sufrido varios cambios profundos en su redacción a lo largo de los años en los que tuvo mayor difusión. La versión contenida en HH1 se considera la primera de estas versiones por varias características fundamentales que la diferencian de los otros testimonios conservados. Rodríguez Puértolas (1968, pp. 105-107) cita, entre ellas, la confusión entre los episodios de la *Presentación en el Templo* y la *Circuncisión*, ya que aparecen unidos en un único pasaje[215]; los ataques a los grandes, esto es, contra Enrique IV, el arzobispo de Toledo y otros personajes cercanos a la figura del rey; la crítica a los dominicos por no aceptar la visión franciscana de la Inmaculada Concepción de María; y la ausencia del episodio de la *huida a Egipto*. Asimismo, identifica como esta primera versión la recogida en los cancioneros manuscritos PN11 (Bibliothèque nationale de France, Esp. 305) y LB3 o *Cancionero de Egerton*[216] (British Library, Egerton 939), dado que mantienen la misma secuencia de capítulos.

No obstante, Whinnom (1977, p. 133) y Severin (2007, p. 225) ponen en entredicho esta clasificación, al considerar la versión de PN11 y LB3 un estadio posterior, esto es, una nueva versión con las coplas contra la nobleza suprimidas —que no sustituidas— del texto principal, pese a que en el testimonio de París están copiadas al final del poema. En efecto, en la redacción principal de estos manuscritos no aparecen las estrofas con los ataques a los grandes, entre otras. Además, PN11 se llega a considerar una copia de LB3 en la que el copista, con toda probabilidad, tuvo acceso posteriormente a otra versión más completa de la obra que contenía los grupos

213 Uhagón identifica 76 poemas y Dutton 97. Garcia atribuye esta discrepancia, respecto a los 92 identificados por él, a errores y diferencias de criterio en el recuento (1990, p. IX).

214 La *Vita Christi* ocupa desde el f. 314^{r} hasta el 345^{v}.

215 Aunque en el evangelio según Lucas aparecen los versículos de la *Circuncisión* y la *Presentación en el Templo*, uno a continuación del otro (Lc 2,21-22), tanto el Levítico como el propio evangelio según Lucas indican que se debe circuncidar al niño a los ocho días de su nacimiento (Lv 12,3; Lc 2,21) y se le presentará al sacerdote cuando se cumplan los días que marca la Ley de Moisés (Lc 2,22) que, en el caso de ser un varón, son treinta y tres después de la circuncisión, es decir, cuarenta días después del nacimiento (Lv 12-4).

216 Para la edición de este cancionero, véase Severin, 2000.

de estrofas eliminados, que incorporó al final con una nota indicando su localización correcta (Rodríguez Puértolas, 1968, pp. 101-105).

Severin (2007, p. 226) añade a esta segunda versión de Whinnom y primera de Rodríguez Puértolas tres copias adicionales que, aunque ambos las mencionan, no las clasifican al ser fragmentarias: la de NH2 o *Cancionero de Vindel* (Hispanic Society of America, ms. B2280), que es una copia errática y parcial de estrofas carente de rúbricas, y las identificadas como SA4b y SA4c, transmitidas por el complejo cancionero SA4 (Biblioteca General Histórica de la Universidad de Salamanca, ms. 2139)[217]. En ninguno de estos tres testimonios fragmentarios aparecen los versos contra el rey de Castilla y sus seguidores, aunque por motivos diferentes. Si bien es indudable que las estrofas contra los grandes están omitidas en NH2, LB3 y PN11, es bien distinto lo que ocurre en SA4b, ya que la última estrofa del folio 13^{v} es justo la anterior al grupo de coplas que contienen los mencionados ataques, que, siguiendo la secuencia de HH1, deberían comenzar en el folio 14^{r}. Sin embargo, a pesar de tratarse de la misma mano, hay un claro cambio de *ductus* y, según parece, un salto en la secuencia narrativa. Aunque este hecho nos podría llevar a pensar en una pérdida de hojas del cuaderno, la descripción material del códice realizada por Rodríguez Ferrer (2007, p. 25) descartaría esta posibilidad. A pesar de ello, este testimonio presenta unas características únicas en su redacción con coplas que no aparecen en ningún otro, lo que bien podría indicar que no se trata de un problema de transmisión textual, sino de una intervención consciente del copista[218]. El último de los tres fragmentos añadidos por Severin a este grupo, identificado como SA4c, contiene únicamente las estrofas dirigidas contra los excesos de los caballeros y del maestre de Calatrava que aparecen en HH1 y no en LB3, por lo que, siendo tan reducido textualmente, es poco productivo hablar de una eliminación consciente del ataque a los grandes, sino que aquí estaríamos ante una cuestión de transmisión textual muy diferente. Precisamente por la brevedad del fragmento, Dutton ni siquiera lo reconoció como parte de las *Coplas de la vita Christi* (ID 0269) y le dio un identificador diferente (ID 4687), al considerarlo un poema anónimo del que el único testimonio habría sido este cancionero, bajo el título *Aplica y enxempla en el thener contrario y mal despender y poco aprouechar con los caualleros*. A este estadio textual de la obra y al correspondiente grupo de testimonios habría pertenecido un manuscrito "en poder, según

217 Para una descripción de este cancionero, véase Rodríguez Ferrer, 2007.

218 Para una transcripción de las estrofas que únicamente aparecen en SA4b, véase Severin, 2017, pp. 293-295.

parece, del Sr. Eugenio Montes" (Rodríguez Puértolas, 1968, p. 88) en la década de los 60 y hoy perdido, que contenía las coplas dedicadas a la reprobación de los vicios que también recoge SA4c, pero con nombres específicos de personajes de la nobleza, una característica única que podría situar este testimonio como antígrafo de HH1.

La tercera versión de Whinnom y segunda de Rodríguez Puértolas sería la que se encuentra en el cancionero EM6, custodiado por la *Real Biblioteca del Monasterio de San Lorenzo de El Escorial* con la signatura K-III-7[219]. En este caso, "representa una etapa entre los MSS. de la versión primitiva y la versión final, corregida, de los textos impresos" (Whinnom, 1961, p. 163). Destaca por suprimirse o suavizarse la mayor parte de las estrofas con ataques políticos y sociales, como por ejemplo las dedicadas a atacar a los grandes, que se cambian por una disculpa, a diferencia de PN11 y LB3, en las que se eliminan directamente, como hemos visto. Severin, adicionalmente, asocia a esta recensión la tercera copia contenida en SA4, la que Dutton designó como SA4a:

> Esta es la versión más larga del manuscrito revisado, que circulaba en la corte al final del siglo xv. Incorpora partes reescritas por extenso, con la mayoría del comentario político expurgado. Sin embargo, no fue la versión utilizada para la imprenta. Hay otra versión de la misma en Salamanca 2139, fols 71r-122r. A pesar de la mucha reescritura, contiene tantas estrofas como HH1. Sus dos últimas estrofas, mal consideradas únicas, también figuran como las dos primeras estrofas de ID 7815, fol. 119r (Severin, 2017, p. 292).

Además de la supresión de las confrontaciones políticas, SA4a y EM6 destacan por modificar la secuencia episódica primitiva para hacerla más acorde con la narración bíblica. Separan la *Circuncisión* de la *Presentación en el Templo* y establecen el siguiente orden: *Natividad*, *Circuncisión*, *Historia de los Reyes*, *Presentación en el Templo*, *Huida a Egipto* y, por último, *Historia de los Inocentes*, que está inacabada, al igual que en las ediciones impresas, y que, por tanto, no finaliza con la oración en nombre de Juana de Cartagena recogida por el resto de testimonios.

Aunque actualmente se encuentra en paradero desconocido, el hoy fragmentado *Cancionero de Barrantes*, que Dutton cataloga como el cancionero

---

219 Para una descripción de EM6, véase Zarco Cuevas, 1926, II, pp. 175-184; Whinnom, 1961; Martos, 2021.

manuscrito perdido ZZ3 (Dutton, 1990-1991, IV, pp. 378-381), también transmitía un testimonio manuscrito de las *Coplas de la vita Christi* anterior a la versión impresa (ZZ3-47)[220]. Así lo atestigua el índice conservado en la Biblioteca de la Real Academia Española, en el fondo que perteneció a Rodríguez Moñino (RM CAJA 98-30), fechado entre 1456 y 1480, cuyo contenido textual está intacto, aunque la hoja se encuentra deteriorada y partida por la mitad en su eje vertical, sin duda por haberse doblado por el blanco que separa las dos columnas de escritura. En el f. 1^{v} de este índice, se recogen las *Coplas de la vita Christi*, que aparecerían justo después de uno de los fragmentos recuperados del *Cancionero de Barrantes*, hoy conservado en la RAE, y también procedente de la biblioteca de Rodríguez Moñino (RM 73)[221].

El cancionero SA5, conservado es esta misma biblioteca (ms. 2244), incorpora entre las obras de Ausiàs March y Juan de Mena, f. 158^{v}, la primera estrofa completa del poema en lo que parece ser una *probatio calami* de difícil filiación ecdótica dada su escasa longitud. También fragmentaria sería la copia de BC3 (Barcelona, Biblioteca de Catalunya, ms. 1967), cuyos ff. 97^{v}-98^{v} contienen las primeras 21 coplas de esta obra, copiadas por una mano diferente a la del resto del códice. Ni Whinnom (1961, 1962 y 1977), ni Severin (2004 y 2007) recogen el testimonio parcial de BC3, ni otro que, en este caso, sí que contempla Rodríguez Puértolas (1968, p. 92), con solo tres coplas (las estrofas 52-54 de su edición), que edita de manera sinóptica por su notable singularidad y divergencia textual (1968, pp. 316-317): se trata del cancionero manuscrito que Dutton identificó después como MN19 (BNE, ms. 4114, ff. 450^{r-v}).

El estadio textual de la tradición impresa de las *Coplas de la vita Christi* es el último de ellos y corresponde a su cuarta recensión, según la clasificación de Whinnom, y a la tercera de Rodríguez Puértolas, a la que, según Dorothy Severin (2007, p. 292) pertenecería también el cancionero manuscrito SA9 (Biblioteca General Histórica de la Universidad de Salamanca, ms. 2762), sin justificarlo textualmente. Se inicia cronológicamente con la edición a la que pertenece el ejemplar único conservado en la *Biblioteca Nacional de España*, con la signatura INC/2159[222], y clasificado por Dutton como 82IM.

---

220 Faulhaber, sin embargo, apunta a que la *Vita Christi* del *Cancionero de Barrantes*, en caso de existir, estaría en otro manuscrito diferente del identificado por Dutton como ZZ3 (BETA manid 2723).

221 Y del que existe una copia del siglo XIX, a manos de Juan Facundo Riaño, conservada en la Hispanic Society of America (Ms. 0173).

222 "Ejemplar mútilo de las hojas signaturas e1 y e8, ésta presumiblemente en blanco. Algunas hojas restauradas burdamente" (Martín Abad, 2010, p. 540, nº M-102).

Según atestigua su colofón, salió de las prensas zamoranas de Centenera el 25 de enero de 1482, y se acompaña del *Sermón trobado* del mismo autor. Está en cuarto y "no existe ninguna otra impresión del siglo xv de esas características de tamaño, del *Vita Christi*" (Pérez Gómez, 1959, p. 36), aunque curiosamente en el periodo post-incunable se volvió a editar en este formato (06VC). Utiliza la letra gótica con el tipo 1*:75G del *Gesamtkatalog der Wiegendrucke* (GW), correspondiente a la tipografía ma14187 del *Typenrepertorium der Wiegendrucke* (TW).

Poco después, Pablo Hurus y Juan Planck imprimieron en Zaragoza *c.* 1982 una nueva edición de esta obra y del *Sermón trobado* (Whinnom, 1962, p. 138), siguiendo el modelo de 82IM, pero en formato *in folio*, a las que añadieron las *Coplas por la muerte de su padre* de Jorge Manrique y el *Regimiento de príncipes* de Gómez Manrique. En esta edición se emplearon los tipos góticos identificados como 1:104G por el GW y ma12011 por el TW, utilizados entre 1481 y 1483 según estas mismas fuentes. De esta edición se conservan tres ejemplares: uno en la *Houghton Library* de la Universidad de Harvard, con signatura Inc. 9509.6 (32.4); otro en la *Real Biblioteca del Monasterio de San Lorenzo de El Escorial*, con signatura X-II-7; y un tercero en la *Biblioteca Comunale* de Palermo, con signatura Esp. XI.F.56 (3), que, aunque parece ser que se encontraba en dos volúmenes separados (Pérez Gómez, 1959, pp. 30-31), hoy está cosido en una única pieza con el orden de las dos últimas obras cambiado.

Hasta la segunda mitad del siglo xx, esta edición zaragozana correspondiente a 82*IM se llegó a considerar la *editio princeps*, fechándola en 1480, anterior a la edición de Centenera.

> El título de *edición príncipe* se encontraba en disputa entre A [82IM], y B [82*IM]. A favor de B, y según González de Amezúa en el Prólogo a la edición facsimilar de A, hecha por la *Real Academia Española* en 1953, se pronunciaba Menéndez Pelayo, y, con él, el Padre Benigno Fernández atribuyéndole la fecha de 1480 (*La Ciudad de Dios*, LVI, 1901) y Augusto Cortina en la edición por él dirigida del *Cancionero de Jorge Manrique* para la colección *Clásicos Castellanos*. (Pérez Gómez, 1959, pp. 32-33)

Sin embargo, no hay duda de que la edición de 82*IM se creó a partir de 82IM, puesto que la primera edición zaragozana de las *Coplas de la vita Christi* reúne los dos impresos poéticos de Zamora (82IM y 82*GM) y le incorpora las *Coplas a la muerte de su padre* de Jorge Manrique, enriqueciendo la edición anterior por criterios comerciales:

> La importancia de este temprano impreso poético no se limita ni a su relación con el primer pliego, ni a la novedad de un modelo de transmisión poética en la imprenta, sino que radica también en lo que debió de ser un arcaico y, sin embargo, paradigmático caso de competencia entre impresores o editores, como otros catalogados en la imprenta incunable valenciana o en la imprenta flamenca de mediados del siglo xvi. Esta circunstancia es la que debió de provocar el enriquecimiento poético de los impresos posteriores de las *Coplas de la Vita Christi* y así, en 82*IM, otro impresor, probablemente en Zaragoza, añadió a esta obra y al *Sermón trobado* la que Centenera había publicado como pliego suelto, no descarto que con anterioridad a las obras de Íñigo de Mendoza: el *Regimiento de príncipes* de Gómez Manrique. La reunión de ambos impresos en uno era el aliciente comercial para una nueva edición frente a la que había salido de las prensas de Centenera. (Martos, 2018b, pp. 527-528)

No solo esta cuestión comercial avalaría esta preeminencia cronológica de la edición zamorana, sino que lo confirmarían también los estudios lingüísticos de estos impresos poéticos salidos de los talleres de Centenera (De Beni, 2024) y de los Hurus (De Beni, en prensa), puesto que estos últimos corrigen buena parte de los rasgos leoneses que caracterizaban la *editio princeps*[223], así como la propia corrección de errores detectada por Whinnom:

> B [82*IM], as I have noted, is a page-for-page reprint of A [82IM], and, though it makes various obvious emendations (*dexemos*, *carrera*, *enla*, etc.) it does not correct a single one of the errors I have listed, though the reader will already have perceived, even without the assistance of the context, how some of them might be remedied. B, as I have also mentioned, omits the six stanzas of A's B3^{r}; but these stanzas are not omitted in any other printed edition. B is, in short, a dead end as far as the *Vita Christi* is concerned, and is useful only for filling the gap of twelve stanzas left by the loss of EI in the sole surviving copy of A. (Whinnom, 1962, p. 145)

Como respuesta a la edición de Zaragoza y en esta misma línea de actuación comercial, Centenera amplió pronto su primera edición, a la

223 Véanse también las descripciones lingüísticas correspondientes en las entradas de estos incunables poéticos del catálogo POECIM (De Beni, 2023a, 2023b).

luz de la salida del taller de Zaragoza, hasta el punto de que desembocó en la creación de un nuevo género editorial con la publicación *c.* 1483 del primer cancionero impreso estrictamente castellano[224]:

> Lo que pudo ser la respuesta comercial de Centenera a ese impreso zaragozano en 83*IM es lo que habría generado, circunstancialmente y por motivos comerciales, el nacimiento del primer impreso castellano, que, a las *Coplas de la Vita Christi* y al *Sermón trovado*, añadía otros dieciséis poemas, muchos de fray Íñigo, pero dando cabida a otros de carácter profano, entre los que se encontraban algunos de Jorge Manrique o de Juan de Mena, cuya obra se difundía aquí por primera vez en letras de molde. (Martos, 2018b, p. 528)

Así, junto a la *Vita Christi*, en 83*IM el impresor zamorano incorporó trece obras más de diversa autoría, la mayor parte del propio Íñigo de Mendoza, y el resto de Jorge Manrique, de Gómez Manrique y de Juan de Mena. Está impreso en folio, con la tipografía 2:93/94G del GW, correspondiente a la ma08073 del TW, que utilizó la oficina de Centenera entre 1483 y 1494. Se conocen cuatro ejemplares, uno en la *Real Biblioteca del Monasterio de El Escorial* (38-I-27), otro en la *British Library* (IB.52920), un tercero en posesión de la Biblioteca Nacional de España (INC/897)[225] y un último en posesión de manos privadas o perdido[226]. Whinnom (1962, p. 137) apunta que Foulché-Delbosc utilizó una copia de este testimonio para hacer la transcripción de las *Coplas de la vita Christi* (1912, pp. 1-52) con la que abre su *Cancionero Castellano del siglo* xv[227].

El éxito de este cancionero incunable (83*IM) y su fortuna comercial acabó repercutiendo editorialmente en sentidos muy diferentes: destaca,

---

224 Más allá del precedente de *Les trobes en lahors de la Verge Maria* (74*LV), como hemos visto.

225 "Ejemplar mútilo de las hojas signaturas g3, g8, y las dos últimas hojas sin signatura, teniendo duplicada la hoja signatura g2, con algunas manchas y restaurado" (Martín Abad, 2010, p. 541, nº M-103).

226 Del que nos da noticia Palau y Dulcet: "En 1951 vimos un ejemplar en poder del Sr. Rodríguez Urtueta de San Sebastián, procedente de la Librería Babra, por el que pedía 40.000 pts. Ignoramos si fue vendido" (1948-1977, IX, p. 40, nº 163763).

227 Lo único que Foulché-Delbosc menciona sobre las fuentes que empleó es que "de las copias que nos han servido para la presente colección, hay algunas que han sido hechas harto lejos de nosotros y que no nos ha sido posible cotejar con sus respectivos originales" (1912, p. VIII).

especialmente, que esta edición del llamado *Cancionero de Centenera* o *Cancionero de Íñigo de Mendoza* fue copiada por Fadrique de Basilea, teniendo en cuenta que los privilegios de impresión, ni mucho menos los derechos de autor, como concepto posterior, eran todavía presente en estos impresos incunables[228]. Se trata de la edición que se recoge en el ejemplar único conservado en la *Library of Congress* de Washington, muy falto de hojas y que Goff identificó erróneamente como un ejemplar del cancionero 95VC, al que nos referiremos después: "Zaragoza: Paul Hurus, 10 Oct. 1495. Fº. Ref.: HR 4313; Haeb(BI); Sánchez 50; Vindel(A) IV 216; Jurz 265. Cop: LC(-)" (Goff, 1973, M-489). Hay una nota inicial en el ejemplar, de hecho, que lo data en 1495, a partir de la primera versión de la *Bibliotheca Hispana* (1672, II, p. 290) de Nicolás Antonio, y que podría haber confundido a Goff, un error que ya corrigieron Rivera y Trienens:

> from its position relative to a manuscript list of contents occupying the same flyleaf, it is clear that this note was a later addition. In fact, its last lines are crowded together so that it can fit above the caption to the list. That the note was written as early as the eighteenth century is suggested by the fact that its page references relate to Nicolás Antonio's *Bibliotheca Hispana* as originally published at Rome in 1672 instead of to the revised edition published at Madrid in 1783 under the title *Bibliotheca Hispana Nova*. (1979, p. 22)

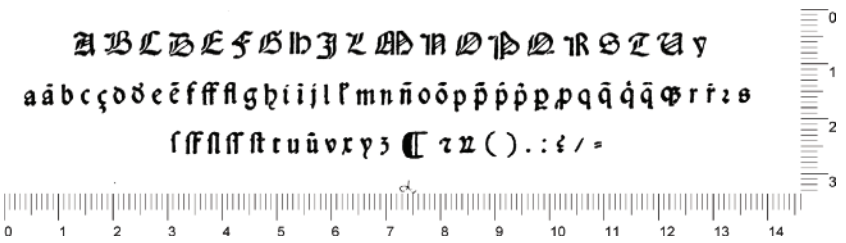
Se ha sabido poco de este impreso en el contexto de los estudios de cancionero, no solo por la confusión bibliográfica, sino por el desconocimiento que tenía de él Dutton, que repercutió en su falta de catalogación. Es por ello que Manuel Moreno, al tener noticia bibliográfica a través de Rivera y Trienens (1979) y necesitar citarlo como testimonio de ciertos poemas, decide asignarle un identificador: "90*IM (No figura en Dutton, propongo la sigla a la manera de Dutton)" (Moreno, 2011, p. 53), una

228 En una época en la que no se habían establecido los derechos de autor, la copia del trabajo ajeno no se consideraba delito, por lo que, si la competencia estaba teniendo éxito en la venta de un producto, el mecanismo de réplica mejorada auguraba un futuro comercial más prometedor que la búsqueda de uno nuevo. Sobre 1480 aparece el *privilegio* de impresión de libros, que permitía al promotor de la edición, a petición propia, imprimir y comercializar una obra con un precio de venta estipulado, la *tasa*. El primer libro al que se le otorgó fue la *Cura de la piedra y dolor de la yjada y colica renal*, de Julián Gutiérrez, médico real, e impreso en Toledo por Pedro Hagenbach en 1498 (Pérez Priego, 2011, p. 92).

referencia que ha tenido cierto éxito en la bibliografía posterior. La datación de este incunable *c.* 1490 no es del todo desacertada y, siguiendo criterios tipográficos, el GW (M1872310) lo data *c.* 1491, mientras que el ISTC (im00489000) y BETA (manid 3646) ofrecen una horquilla más amplia (1490-1493), en todos los casos aceptando ya que es un producto del taller burgalés de Fadrique de Basilea. Fue impreso en folio con el conjunto de tipos clasificado como 7:97G en GW e identificado como ma03789 por TW, que se utilizaron en este taller entre 1490 y 1500. Es curioso destacar de esta tipografía que comparte rasgos de diseño con muchos de los tipos utilizados por Pablo Hurus en su *Cancionero de Zaragoza* en 1492 y 1495, aunque en este caso ya es una 100G (Figura 16) y no una 97G (Figura 15), si bien la *M*, precisamente, responde a un diseño muy diferente. Es, quizás, tal parecido, junto a la singularidad de ejemplares y su dispersión, así como el carácter mútilo de 90*IM, lo que complicó la adscripción tipográfica de este ejemplar.

Figura 15. Muestra tipográfica de 90*IM (Typenrepertorium der Wiegendrucke, ma03789)

Figura 16. Muestra tipográfica de 92VC/95VC (Typenrepertorium der Wiegendrucke, ma03787)



El cancionero 83*IM, en cualquier caso, era la fuente directa o indirecta para 90*IM y 92VC/95VC[229], aunque en términos muy diferentes, pues uno es una copia fiel, mientras que el otro es el modelo a imitar en la creación de un nuevo cancionero de fray Íñigo: el conocido como *Cancionero de Zaragoza* (92VC/95VC). En efecto, el taller de los Hurus fue mucho más creativo que el burgalés de Fadrique de Basilea, especialmente en lo que respecta a la reproducción de este cancionero impreso de Centenera. El taller zaragozano siguió una cuádruple estrategia ante la gran novedad comercial que supuso la impresión de 83*IM: a) la impresión de un nuevo cancionero que compitiera con el de Centenera, pero, en este caso, centrado en otro gran autor, como era Fernán Pérez de Guzmán, con lo que se gestó el *Cancionero de Llabià* (86*RL) como nuevo producto poético[230]; b) la impresión de nuevas ediciones de las *Coplas de la vita Christi* de fray Íñigo (91*VC); c) nuevos impresos de poesía, con obras inéditas hasta entonces y con cierto carácter antológico, como la *Pasión trobada* (91*PT) que ha localizado recientemente Fernández Valladares (2019); y d) la creación de un nuevo cancionero de fray Íñigo, casi una década después, que superara, desde diferentes perspectivas, el de Centenera, con una eclosión de nuevos textos y, sobre todo, de xilografías, que lo enriquecían y ennoblecían (92VC y 95VC).

En el ejemplar incunable del *Cancionero de Zaragoza* (92VC) adquirido recientemente por la Biblioteca Nacional de España con la signatura INC/2900, se encuentra interpolado un amplio fragmento de una edición zaragozana de la *Coplas de la vita Christi*, al que Fernández Valladares ha dado el identificador 91*VC, siguiendo el modelo de Dutton, al encontrarse

---

229 "An even closer dating is suggested by a textual comparison of *Vita Christi*, *Razón y sensualidad*, *Gozos*, *La cena*, and *La Verónica* as printed in the Library of Congress edition (C2) and in the *Cancioneros* of 1483-84? (C1) and 1495 (D2). The *Vita Christi* text in D2 appears to derive not from C1, as heretofore believed, but from C2 through the now lost Paul Hurus edition of 1492 (D1), D2 being "a page-for-page, line-for-line reprint of D1" (Rivera & Trienens, 1979, p. 25).

230 "Eso no significa, sin embargo, que desde Zaragoza no hubiese habido anteriormente ninguna reacción o secuela a este cancionero de Centenera, por bien que no inmediata, porque creo que hay que interpretar en este sentido el *Cancionero de Ramón de Llabia* (86*RL), a la luz del tipo de obras incorporadas, aunque Fernán Pérez de Guzmán sustituya el protagonismo de fray Íñigo de Mendoza, puesto que la composición y el tono de ambos productos es más que similar" (Martos, 2018b, pp. 528-529). Para este cancionero impreso del taller de los Hurus, véase López Casas, 2020, 2021; López Casas & Mangas, 2022; Martos, 2024c.

ausente, como es lógico, de este catálogo de fuentes poéticas[231]. La ausencia de colofón dificulta su datación, aunque su tipografía, clasificada como 1:99G por el GW y como ma01235 por TW, sitúa su impresión en las prensas zaragozanas de los Hurus a partir de 1485, mientras que es la aparición de 92VC lo que matiza su *terminus ante quem*, pues habría de ser una edición anterior al *Cancionero de Zaragoza*, esto es, impresa entre el año de emergencia de esa tipografía y 1491, tal como apunta Fernández Valladares (2019, pp. 72-73).

El *Cancionero de Zaragoza* es uno de los grandes hitos en cuanto a los impresos poéticos incunables tanto en el contexto hispánico, como en el propio taller de los Hurus, por tratarse de un nuevo cancionero[232] que, además, incorpora ricas estampaciones xilográficas, descritas por Lacarra (2020, pp. 115-122), con un total de "60 estampas del ancho de columna pertenecientes la mayoría a dos series del ciclo de la *Vida y Pasión de Jesucristo*, junto con una de la *Virgen del Pilar*, otra algo más pequeña de *la Verónica* y una a plana entera de la *Muerte portadora de un ataúd y una saeta*. Merece la pena destacar que para 33 de las 45 entalladuras diferentes esta es la primera aparición documentada" (Fernández Valladares, 2019, p. 58). La estructura interna de 92VC se abre con las *Coplas de la vita Christi* de fray Íñigo, que inaugura una novedosa secuencia de obras, cuyo cuidado diseño busca dar continuidad a la abrupta terminación de este poema en el episodio de la *Historia de los Inocentes*, incorporando como su continuación natural las *Coplas de la cena* del mismo autor, la *Pasión trobada* de Diego de San Pedro junto a las *Coplas a la Verónica*, también del franciscano, y cerrando el relato de la vida de Jesús con la *Resurrección* de Pedro Jiménez[233].

---

231 "Así, estos dos fragmentos estarían atestiguando sendos cancioneros anteriores, aunque cercanos cronológicamente a 92VC, con cuyo contenido serían —al menos en parte— coincidentes. Según eso, el fragmento de la *Pasión trovada* pasaría a ser el primer testimonio impreso conocido de esta obra (con el código convencional 91*PT) y el fragmento de la *Vita Christi* (91*VC), por su factura claramente distinta de la de 92VC, nos habla de la diversificación editorial de un mismo producto en dos modalidades, una de lujo con estampas —hasta 22 en la de 1492 y otra previa y más barata, sin ellas" (Fernández Valladares, 2019, p. 73).

232 El otro gran cancionero castellano incunable, unido a 83*IM [=90*IM], 86*RL y 92VC/95VC será *el Cancionero de Juan del Encina* (96JE) (Martos, 2018b, pp. 528-529).

233 "La antología poética se abre con la *Vita Christi* de fray Íñigo, que va desde la Anunciación hasta la matanza de los Inocentes. En los poemas que vienen después, el ideal relato de vida de Jesús llega directamente a sus momentos finales, pues se suceden cuatro obras sobre su muerte y resurrección, una de ellas, la tercera, dedicada a la Verónica" (Marini, 2023a).

Se había creído que todos los ejemplares de esta edición estaban perdidos hasta el trabajo de Fernández Valladares (2019), que da noticia del adquirido por esas fechas por la Biblioteca Nacional de Madrid (INC/2900)[234], consultado y estudiado por ella con anterioridad a su llegada a este fondo, así como demuestra que el de la Bibliothèque de l'École nationale supérieure des Beaux-Arts de París (Masson 2055) pertenece también a 92VC y no, como se había creído hasta entonces, a 95VC, del que solo se conserva, por tanto, el de la Biblioteca Alessandrina, como veremos. Tanto una edición como otra presentan colofón impreso que atestigua el taller concreto, como lo hace su propia tipografía: se utiliza para los textos poéticos la 3*100G, según el GW, identificada como ma01247 por el TW, que empleó este taller desde 1490 hasta 1499, mientras que, para los titulillos y rúbricas, se emplea la 2*134G del GW, ma01246 del TW.

Muestra del éxito que debió de tener la que debió de ser la *editio princeps* del *Cancionero de Zaragoza* es que, solo tres años después, en 1495, se hiciese una nueva edición que, prácticamente, es copia a plana y renglón, compartiendo ilustraciones, de 95VC[235], impresa también por Pablo Hurus, como atestigua su colofón y la propia tipografía. Está impresa en folio, al igual que el resto de ediciones incunables a excepción de la *princeps*, con las mismas tipografías utilizadas por 92VC. Como hemos avanzado, únicamente se conserva un ejemplar en la Biblioteca Universitaria Alessandrina de Roma (Inc. 382), del que ya acusa noticia en esta misma localización Nicolás Antonio en su *Bibliotheca Hispana Nova* (1783, I, p. 361), si bien se había creído hasta 2019 que el de París de 92VC lo era también.

---

234 La coincidencia de la signatura con 91*VC se debe a que ambos comparten el mismo volumen. Para superarlo, BETA propone distinguir las tres unidades de la siguiente manera: 92VC como INC/2900(1) (manid 2005); 91*PT como INC/2900(2) (manid 6191); 92VC como INC/2900(3) (manid 6192). Para una descripción de las respectivas ediciones y de los tres ejemplares que se conservan del *Cancionero de Zaragoza* (92VC y 95VC), véase Fernández Valladares, 2019; Marini, 2023a, 2023b, 2024, en prensa.

235 La principal diferencia entre una edición y otra es esta que señala Marini: "El ejemplar de 95VC omite deliberadamente el grabado número 60, que en 92VC se halla en el centro de la página, al principio del anónimo 'Dezir de la muerte' (texto nº 15), desprovisto de rúbrica, con solo el titulillo, que por otro lado comparte con el poema subsiguiente, otro 'Dezir de la muerte', esta vez de Fernán Pérez de Guzmán (texto nº 16). En 95VC no solo desaparece el grabado con la Muerte, sino que también se añade una rúbrica antes del íncipit del texto que comienza en el f. CXII^{v}: 'Un dezir gracioso de la Muerte', que sin embargo no aparece en la tabla al principio del incunable, tal como ocurría también en 92VC" (2024, p. 196).

Son estas ediciones de 92VC y/o 95VC el origen de que se incorporen dos manuscritos a los testimonios impresos de la última recensión de las *Coplas de la vita Christi*, ya que ambos son *codices descripti*, esto es, copias manuscritas fieles elaboradas a partir de un impreso copiado, a manera de antígrafo[236]. El cancionero MN46 (Biblioteca Nacional de España, MSS/18183) lo es de la *editio princeps* (92VC) (Díez Garretas, 2010)[237], mientras que no queda clara la fuente de ML1 (Fundación Lázaro Galdiano, MS30), aunque, sin duda, es una de estas dos ediciones, entre las que Díez Garretas (2011, pp. 84-95) apunta a 95VC[238], sin argumentos que lo confirmen. Sea como fuera, ambos testimonios manuscritos son copias de impresos y, como tales, no tienen valor ecdótico, si no es que ha habido algún tipo de contaminación de la tradición, por recurrencia a otras fuentes, como llega a sugerir Severin[239].

La última edición incunable conocida de las *Coplas de la vita Christi* se imprimió en Sevilla en 1499 por Meinardo Ungut y Estanislao Polono por mandato de Lázaro de Gazanis, esto es, en su función de editor o patrocinador, tal como indica su colofón. De ella únicamente se conserva un ejemplar mútilo que forma parte de la colección de The Morgan Library & Museum de Nueva York, con la signatura PML 76441, y que posiblemente sea el mismo que sitúa Vindel en una librería de Barcelona en

---

236 Para este mecanismo de transmisión es esencial la monografía que coordina Josep Lluís Martos (2011a) y los trabajos incorporados a ella, en especial sus conclusiones (Martos, 2011b), por su perspectiva general, así como su estudio particular de un *codex descriptus* de Ausiàs March (Martos, 2014).

237 Fernández Valladares propone, incluso, una convincente hipótesis en cuanto a su autoría: "Ello lleva a pensar que el manuscrito MN46 pudiera ser, precisamente, la copia sacada por Vázquez Espina para Méndez, porque en él aparece transcrita esa nota de procedencia en la última hoja y además en disposición invertida, lo que pudo motivar que pasara desapercibida para Méndez. Pero el detalle que nos parece más concluyente es la cuidadosa escritura de esa copia y, en particular, la perfección con que aparece dibujada a plumilla la marca tipográfica de Hurus que figura en el impreso bajo el colofón" (2019, p. 64, n. 29).

238 "El contenido de la tabla de su cancionero se corresponde con dos impresos editados en Zaragoza en 1492 (E) y en 1495 (F), en las prensas de Pablo Hurus. Dos impresos aparentemente iguales y, según se venía diciendo, el segundo probablemente reedición del primero, por lo que cualquiera de los dos pudo servirle de copia. El cotejo de ambos, aunque del primero sólo se conservan las tres últimas composiciones, manuscritas en el siglo XVIII (BNE, ms. 18183), demuestra que se trata de dos ediciones distintas, y de ellas nuestro amanuense se sirvió de la más cercana a la fecha de su copia, la de 1495 (F)" (Díez Garretas, 2011, pp. 84-85).

239 "Curiously, it has small number of stanzas from the original tradition (HH1) missing form the Third recension (SA4a, EM6)" (Severin, 2007, pp. 226).

los años 30 (1945-1954, V, pp. 314-316, nº 117 *Sevilla*)[240]. Está impreso en folio con diversas tipografías: para las coplas se utiliza la 5:98G de estos impresores según el GW, correspondiente a ma01526 según el TW, mientras que las rúbricas se imprimen con la 2:112G del GW (TW ma06455) y los titulillos y parcialmente alguna rúbrica lo hacen con la 4:150G del GW (TW ma01525). Aunque la edición se enriquece, como en el caso de 92VC y 95VC con xilografías, no son tantas en esta edición sevillana como en el *Cancionero de Zaragoza*.

Tenemos noticias, al menos, de dos ediciones post-incunables, la primera de las cuales, desconocida por Dutton, es un fragmento de ocho hojas de la *Vita Christi* adosado al final de un ejemplar mútilo de las *Coplas de Bias contra Fortuna*, del marqués de Santillana, impreso en cuarto en Toledo (BNE, R/12340), y que Norton atribuye a Polono en Sevilla *c.* 1502 (1978, pp. 278-279, nº 740). La otra edición post-incunable, también en cuarto, de la que se conserva igualmente un ejemplar único en la Biblioteca Nacional de España (R/11897), aunque en este caso completo, se imprimió en Sevilla en el taller de Jacobo Cromberger en 1506, con lo que Dutton la catalogó como 06VC. Como venía siendo habitual desde la *editio princeps* del *Cancionero de Zaragoza* (92VC), ambas ediciones post-incunables de las *Coplas de la vita Christi* presenta grabados xilográficos y parecen ser copias de 99VC (Whinnom, 1962, pp. 143-144).

Es muy probable que una edición de mediados del siglo xvi de esta obra, que salió de las prensas de Jacobo Cromberger en 1546, se confeccionase a partir del post-incunable sevillano de 1506. Esta edición tardía se conserva también en ejemplar único, de nuevo en cuarto y en la BNE (R/12775). Del siglo xvii, se tienen noticias de dos ediciones más con el título *La vida de Christo en metro Castellano*: una sevillana de 1611 en octavo y otra vallisoletana de 1615. Nicolás Antonio (1783, p. 361) las cita refiriendo el *Catálogo de libros hispanos* de Tomás Tamayo, aunque desafortunadamente a día de hoy no existen referencias de ningún ejemplar accesible asociado a alguna de ellas.

Ya en el siglo xix, dentro del gran proyecto ilustrado del *Cancionero general del siglo xv* (Martos, 2012a), avalado por Carlos IV en 1807 y recogido en once volúmenes manuscritos, que Dutton cataloga como

240 "Desconocido a Haebler. No tenemos más referencias de este libro que las presentes reproducciones, que fueron obtenidas en 1930 de un ejemplar falto que se conservaba en la librería de D. Salvador Babra, de Barcelona, ignorando en la actualidad su paradero" (Vindel, 1945-1954, V, pp. 314-316, nº 117 *Sevilla*).

MN13 (Biblioteca Nacional de España, mss. 3755-3765), figura, en su tercer volumen, una copia completa de las *Coplas* del franciscano que abarca desde el f. 258r al f. 395r (Díez Garretas, Martos & Moreno, 2012). Según apunta Moreno (2012b, p. 21), esta copia manuscrita decimonónica de las *Coplas de la vita Christi* se hace a partir de su *editio princeps* (82IM), tal y como indica el propio cancionero manuscrito.

Las *Coplas de la vita Christi* de fray Íñigo de Mendoza no se volverán a editar hasta principios del siglo xx, de mano de Foulché-Delbosc, que las transcribe a partir de 83*IM para inaugurar su *Cancionero Castellano del siglo xv* (1912, I, pp. 1-52). En 1968, Rodríguez Puértolas publicaba dos ediciones críticas. La primera de ellas (1968a) es una edición sinóptica parcial con la *editio princeps* (82IM) en la primera columna, mientras que en la segunda ofrece una versión que reúne los dos estadios textuales anteriores a partir de LB3, PN11 y EM6, sin considerar todavía HH1, al que no debió de tener acceso por estar en ese momento en manos privadas[241]. Su aparato crítico se limita a los testimonios impresos conocidos entonces, así como a lecciones divergentes de los cancioneros manuscritos EM6, LB3 y PN11[242]. En su segunda edición (1968b) optó por transcribir la obra de fray Íñigo de Mendoza tomando como texto base la *editio princeps* (82IM) e incluyendo las variantes al final de la obra. Massoli, en 1977, también optó por realizar una edición crítica del poema a partir de 82IM (1977, pp. 119-225), de nuevo sin referencias al *Cancionero de Oñate*, con las variantes en el aparato, y acompañada de su traducción al italiano[243].

---

241 "Su actual poseedor es el profesor Edwin Binney, de la Universidad de Harvard, quien se ha negado, como señalo en nota preliminar, a permitirme el estudio" (Rodríguez Puértolas, 1968a, p. 85).

242 "En la segunda parte, la edición crítica del poema de Mendoza, coloco la versión A (es decir, la tercera impresa) en la primera columna; la de b1 (segunda versión) y a1 y a2 (primera versión), en la segunda columna. En la transcripción de la primera versión, un guion (—) al pie de una copla y ante la sigla a1 indica variante en este texto de París, así como que la lectura de la copla a que se refiere es ecléctica entre a1 y a2. Puntos suspensivos (…) en un verso de la primera versión indica que la continuación del mismo es exacta en la copla equivalente de la posterior versión. En las notas a pie de página, un asterisco (*) indica importante detalle textual. Dos o más asteriscos suponen la existencia de dos o más detalles importantes. Un asterisco junto al número de copla indica que hay comentario a la misma en las notas al texto" (Rodríguez Puértolas, 1968a, pp. 9-10).

243 Para la caracterización de estas tres ediciones, véanse los apartados correspondientes de las fichas POECIM/92VC y POECIM/95VC del catálogo *Poesía, Ecdótica e Imprenta* (Marini, 2023a, 2023b).

Contamos, asimismo, con dos ediciones facsimilares: a mediados de siglo la reprodujo en facsímil la *Real Academia Española*, a partir del ejemplar único de la edición de 82IM (Mendoza, 1953), al que, como hemos visto, le falta la h. e[j], circunstancia que se resolvió mediante una composición fotográfica para acomodar las estrofas faltantes a partir de la edición de 83*IM, como advierte Pérez Gómez (1959, p. 36), quien, por esto mismo, ofreció en 1975 un nuevo facsímil, pero, en este caso, a partir del ejemplar de Palermo de 82*IM (Mendoza, 1975).

Este recorrido por los numerosos testimonios de las *Coplas de la vita Christi* nos confirma que su época de esplendor fue la segunda mitad del siglo xv, no solo por el número de manuscritos que se han conservado, sino por la cantidad de ediciones que salieron de los talleres durante este periodo, enumeradas de forma sinóptica en la tabla adjunta.

Como ya se ha argumentado previamente, la creación de un modelo para una inteligencia artificial que sea capaz de reconocer el texto contenido en una imagen pasa por suministrarle la asociación entre la grafía y su correspondiente letra, el proceso que llamamos *entrenamiento*. En el caso de emplear una sola tipografía en este proceso, únicamente será capaz de transcribir documentos creados con ella. Actualmente, la homogeneización tipográfica producto de la utilización de los ordenadores ha facilitado la construcción de modelos capaces de transcribir con tasas de éxito cercanas a la perfección en obras contemporáneas. Estos modelos suelen estar entrenados con las familias tipográficas más difundidas: *Verdana*, *Arial*, *Times New Roman*, entre otras. No obstante, la elaboración artesanal de los tipos medievales limita esta homogeneización. Cada taller impresor disponía de varias familias tipográficas únicas que utilizaba durante un periodo de tiempo hasta que los punzones, las matrices o, en última instancia, los tipos comprados a un artesano, se desgastaban.

Por tanto, si no se quiere limitar la transcripción a un conjunto de obras de un determinado taller durante un acotado periodo temporal, habrá que utilizar un muestrario representativo de la tipografía empleada durante este periodo. Es por esta razón que, al seleccionar el corpus de estudio, advertimos que las *Coplas de la vita Christi* de fray Íñigo de Mendoza es la obra poética de la que más ediciones incunables conocemos, hasta ocho, que no solo abarcan, prácticamente, todo el periodo del siglo xv en que encontramos poesía de cancionero impresa, desde 1482 hasta 1499, sino que también ofrecen una muestra muy representativa de talleres de imprenta dispersos por la geografía española, todos ellos, además, productores de otros incunables poéticos. A esto habrá que sumar que estos ocho incunables nos ofrecen hasta diez tipografías diferentes, una variedad tal para una misma obra que nos permitirá entrenar un modelo extendido capaz de

Tabla 2. Relación de incunables de las *Coplas de la vita Christi*

Edición	Año	Lugar	Impresor	GW-ID	TW-ID	Ejemplares
82IM	1482	Zamora	Antonio de Centenera	1*:75G	ma14187	Madrid, Biblioteca Nacional de España, INC/2159
82*IM	[1482-1483]	[Zaragoza]	[Pablo Hurus]	1:104G	ma12011	Cambridge, Houghton Library, Inc. 9509.6 (32.4) San Lorenzo de El Escorial, Real Biblioteca del Monasterio de San Lorenzo de El Escorial, X-II-17 Palermo, Biblioteca Comunale, Esp.XI.F.5
83*IM	[1483]	[Zamora]	[Antonio de Centenera]	2:93/94G	ma08073	San Lorenzo de El Escorial, Real Biblioteca del Monasterio de San Lorenzo de El Escorial, 38-I-27 Londres, British Library, IB.52920 Madrid, Biblioteca Nacional de España, INC/897
90*IM	[1485-1491]	[Burgos]	[Fadrique de Basilea]	7:97G	ma03789	Washington, Library of Congress, Incun. X .M52 PQ6180
91*VC	[1485-1490]	[Zaragoza]	[Juan/Pablo Hurus]	1:99G	ma01235	Madrid, Biblioteca Nacional de España, INC/2900
92VC	1492	Zaragoza	Pablo Hurus	3:100G 2*:134G	ma01247 ma01246	Madrid, Biblioteca Nacional de España, INC/2900 París, Bibliothèque de l'École nationale supérieure des Beaux-Arts de París, Masson 2055
95VC	1495	Zaragoza	Pablo Hurus	3:100G 2*:134G	ma01247 ma01246	Roma, Biblioteca Universitaria Alessandrina, Inc. 382
99VC	1499	Sevilla	Meinardo Ungut/Estanislao Polono	5:98G 2:112G 4:150G	ma01526 ma06455 ma01525	Nueva York, The Morgan Library & Museum, ChL f 1734 K [PML 76441]

transcribir no solo otros incunables poéticos, sino, incluso, cualquier otro impreso en gótica de décadas posteriores.

No solo nos ofrece unidad textual la propia elección de una obra concreta, sino que, como hemos visto, los testimonios impresos de las *Coplas de la vita Christi* presentan univocidad ecdótica, esto es, un mismo estadio dentro de su compleja transmisión textual, precisamente porque se utilizaba una edición previa como original de imprenta para la siguiente. Ambos rasgos facilitarán la creación de un modelo extendido para la transcripción automática, pero también permitirán el establecimiento de una edición sinóptica de la cuarta recensión de esta obra, correspondiente a la tradición impresa incunable, como punto de partida para un proyecto general que reúna las fases textuales anteriores, de mayor diversidad y transmitidas por testimonios manuscritos, lo que, en cualquier caso, supera los límites de esta investigación.

En este proceso, la mutilación sufrida por varios de los ejemplares conservados es uno de los problemas a los que hay que enfrentarse, ya que la pérdida parcial de hojas aumenta la complejidad de la comparación textual directa entre las propias ediciones. Dado que se conservan varios ejemplares de algunos testimonios con fragmentos mútilos diferentes, esta transcripción permitirá que en una futura edición se puedan fusionar mediante el desarrollo de algoritmos informáticos para obtener el texto resultante de la combinación de todos ellos.

El corpus que se utilizará para la generación del modelo extendido de transcripción automática queda delimitado, por tanto, por sus peculiaridades y amplitud, por la oportunidad única que ofrece su variación tipográfica a partir de una obra poética concreta de cierta extensión. Ahora bien, habrá que decidir el software a emplear para ello, lo que, obviamente, condiciona todo el trabajo posterior. Aunque la mayoría de investigadores han utilizado la aplicación *Transkribus* en las aplicaciones que se han hecho empleando la inteligencia artificial en el reconocimiento automático de impresos de la literatura castellana medieval y, sobre todo, de los Siglos de Oro, su uso licenciado implica un desembolso económico por cada página transcrita que no hay que obviar. Frente a ello, sin embargo, las alternativas gratuitas no gozan de tanta difusión y, por tanto, aún no se dispone de datos de rendimiento que favorezcan su uso, por lo que, antes de decantarse por un software en concreto y seguir las inercias de la investigación en esta área, es necesario efectuar una comparación en igualdad de condiciones, sin prejuicios adquiridos, que permita su elección en base a los objetivos que se pretenden conseguir.

5.2. Software de transcripción automática aplicado a la poesía de cancionero

Con la proliferación de la inteligencia artificial aplicada al reconocimiento de caracteres, han surgido diversos softwares que permiten el entrenamiento de modelos o que incluso proporcionan modelos ya entrenados, especialmente adaptados a material antiguo. Con el fin de seleccionar el más adecuado a las características de la presente investigación, se han seleccionado tres de ellos, por las razones que se describen después, para efectuar una comparación y evaluar su rendimiento: *Transkribus*, dada su difusión entre los investigadores; *OCR4all*, de la Universidad de Würzburg, al integrar en una misma interfaz todos los procesos y ser de libre distribución; y *eScriptorium*, financiado por varias instituciones francesas, también de libre distribución, y con gran difusión en el mismo ámbito francófono.

5.2.1. Alternativas de software adaptado a material antiguo

Aunque *Transkribus* comenzó como un proyecto colaborativo con acceso libre, pronto se rentabilizó su uso con la creación de la empresa READ-COOP SCE y, aunque sigue disponiendo de una interfaz web que se puede utilizar sin registro, esta tiene una funcionalidad muy limitada. En caso de quererlo utilizar, únicamente se tendrá disponible la opción de enviar una imagen y obtener directamente su transcripción, sin posibilidad de agregar múltiples imágenes, corregir la segmentación, ni crear modelos personalizados. Esta versión es más una demostración del producto con fines comerciales que un software para utilizar en el día a día. La funcionalidad plena la obtendremos después de un registro donde se nos solicitarán nuestros datos y por el que conseguiremos, sin coste alguno, 500 créditos, que nos servirán para empezar a utilizar la aplicación. El coste de uso se aplica de la siguiente manera: de una página manuscrita se nos restará un crédito y de una página impresa, un cuarto de crédito. Cuando se nos acaben estos créditos de bienvenida, podremos renovarlos mediante una suscripción o comprarlos individualmente. Adicionalmente, se nos darán 100 créditos mensuales, aunque con ciertas limitaciones, como el número máximo de modelos que podemos entrenar mensualmente, que no podrán superar los cinco[244]. Aunque no haya que realizar un desembolso inicial, si está previsto darle un uso continuado, inevitablemente se acabará pagando, ya que el volumen de créditos que se conceden sin coste no permite

244 Esta limitación se introducirá a partir del tercer trimestre del 2024.

su aplicación más allá de un reducido conjunto de digitalizaciones. Pese a ello, sigue siendo el software de elección por la comunidad de filólogos del mundo hispano que trabajan con material antiguo. Esta circunstancia lleva aparejada que sus resultados mejoren día tras día con las nuevas aportaciones en forma de modelos que se vienen realizando desde diversos grupos de investigación.

OCR4all es una alternativa gratuita surgida de la tesis de Reul (2020), en la que se partía de la premisa de automatizar el flujo de trabajo que se realiza en un OCR para facilitar su uso por personal no experto en la materia. Para llevarlo a cabo, seleccionó las diversas aplicaciones que había en el mercado que cubrían cada paso de forma individual tras un amplio análisis y las juntó bajo una misma interfaz. Actualmente, el proyecto está colaborando de forma estrecha con *OCR-D*, una alternativa también gratuita y que nació con el objetivo de digitalizar los impresos en alemán publicados entre los siglos XVI y XVIII de forma masiva, con la mínima intervención humana. Así como *OCR4all* surgió como una forma de simplificar el proceso de transcripción, *OCR-D* se centró en su aplicación a gran escala. Con estas visiones complementarias, ambos decidieron en 2020 cooperar para unir esfuerzos, compartir los resultados y no duplicar trabajo, manteniendo cada uno su independencia[245]. Ambos reciben apoyo por parte de varias instituciones alemanas, algo que se refleja en los modelos de transcripción que soportan de base, algunos de ellos centrados en tipografías empleadas mayoritariamente en esta región, como es el caso de la *Fraktur*.

La última de las propuestas, *eScriptorium*, comparte esta misma filosofía de ofrecer una alternativa gratuita y abierta para digitalizar documentos antiguos, pero en este caso, con un uso mayoritario en el ámbito francés. Es un proyecto auspiciado por *eScripta*, el grupo de humanidades digitales de la Université Paris Sciences et Lettres (PSL). Está bajo el paraguas de *Scripta PSL*[246], que engloba, además de *eScriptorium*, varias herramientas de apoyo para la edición y transcripción de textos como *Kraken*, el software

245 En la web de *OCR4all* se anuncia la colaboración. Para más información, acceder a https://www.ocr4all.org/about/ocr4all [consulta: 19/09/2023].

246 Según la página web del proyecto en https://escripta.hypotheses.org/about, "the programme 'Scripta-PSL, The History and Practices of Writing' aims at integrating the fundamental sciences that deal with written artefacts (palaeography, codicology, epigraphy, history of the book, etc.), with other disciplines in the humanities and social sciences (linguistics, history, anthropology, etc.), together with digital and computational humanities, around the study of writing. The program, a Strategic Interdisciplinary Research Initiative (IRIS) of PSL, is supported by EPHE and EFEO, in association with ENS, ENC, EHESS, Collège de France and IRHT (CNRS)" [consulta: 03/05/2024].

que utiliza *OCR4all* para el proceso de segmentación y *eScriptorium* tanto para la segmentación como para la transcripción.

5.2.2. La poesía incunable ante la segmentación y el reconocimiento de grafías

Como ya se ha comentado, a la hora de abordar una comparativa del software para la transcripción de material antiguo cobran especial relevancia dos pasos por sus especiales características: la segmentación y el reconocimiento de grafías. Pese a que ya existen investigaciones que han abordado la transcripción automática de textos del periodo incunable, todas ellas se han basado en la prosa y no se han enfrentado a la principal característica diferencial de la poesía: su puesta en página. La complejidad asociada a este hecho viene derivada de la longitud de cada línea de texto, tanto si la composición es a una como a dos columnas[247], ya que la métrica de los versos no ofrecerá un mismo número de grafías, a pesar de su regularidad de cómputo, como es lógico, pues un pie métrico puede estar formado por solo una o alcanzar las cinco o seis, si no más. A ello se suma que cada grafía tiene un tamaño y, por ejemplo, la letra *m* ocupa más espacio que la letra *i*. A diferencia de la prosa, además, no es posible desplazar palabras o sílabas de una línea a la anterior, lo que supone un inconveniente de composición, si bien es cierto que facilita la cuenta del original. Por todo ello, el texto poético no permite la fluidez del texto en su puesta en página de la prosa, sino que tiene una estructura fija que se debe mantener.

Aunque actualmente la edición por medio de un ordenador permite alinear los márgenes de forma precisa con la inserción de espacios entre palabras y letras de forma automática y casi imperceptible para el ojo humano, el sistema de composición de los tipos móviles de las prensas manuales es rígido y deja poco juego al cajista para ajustarlos a la par. Únicamente la incorporación de piezas ciegas entre cada una de las palabras hubiese permitido, en cierta forma, su ajuste, aunque a expensas de un arduo trabajo que no acababa llevándose a cabo, por considerarse innecesario en la época.

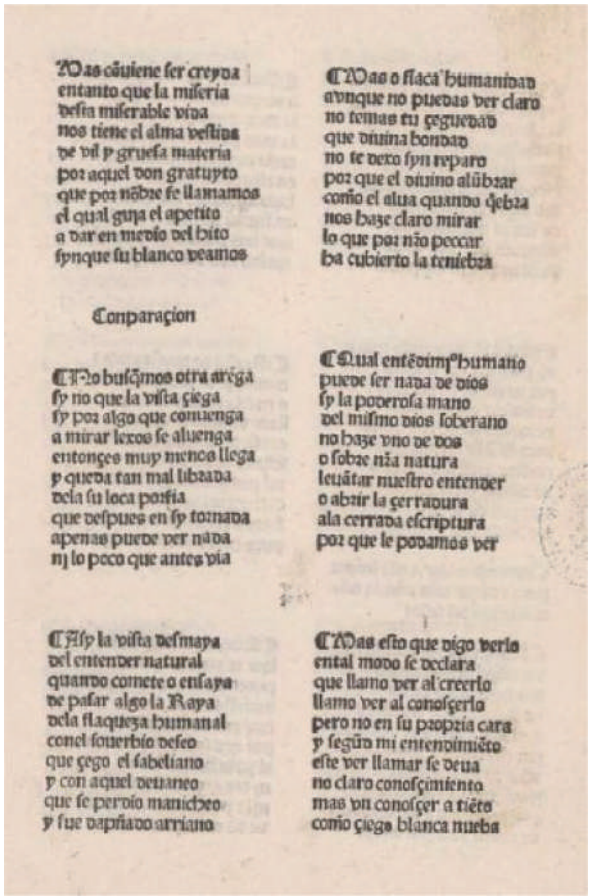
El proceso de segmentación también se va a enfrentar a otro problema añadido en las ediciones de las *Coplas de la vita Christi*, ya considerado en la aplicación de estos OCR a impresos en prosa con tipografía gótica: la interpretación de un texto a dos columnas. Si bien un lector humano habituado a textos con alfabeto latino no tiene ningún problema en identificar su orden de lectura, puesto que se lee el texto completo columna a

247 E, incluso, más columnas en post-incunables poéticos en tipografía gótica, como el *Cancionero General* de Hernando del Castillo (11CG/14CG).

columna, empezando por la que está más a la izquierda, este conocimiento adquirido no es tan sencillo aplicarlo en un algoritmo. A esto se sumarían las diferentes casuísticas generadas durante el proceso de digitalización, casi nunca controlado por el investigador, entre las cuales destaca el descuadre de la alineación producido por la curvatura de la página.

La poesía incunable ofrece otros problemas añadidos en cuanto a su puesta en página, como son los espacios interestróficos en blanco. La Figura 17 muestra cómo la composición de la página busca alinear los

Figura 17. Espacios interestróficos irregulares (Biblioteca Nacional de España, INC/2159, h. a5ʳ – 82IM)


Mas cõuiene ser creyda
entanto que la miseria
desta miserable vida
nos tiene el alma vestida
de vil y gruesa materia
por aquel don gratuyto
que por nõbre se llamamos
el qual guya el apetito
a dar en medio del bito
synque su blanco veamos

Mas o flaca humanidad
avnque no puedas ver claro
no temas tu çeguedad
que diuina bondad
no te dexa syn reparo
por que el diuino alũbrar
cõño el alua quando deja
nos haze claro mirar
lo que por não peccar
ha cubierto la tenieba

Comparaçion

No buscamos otra arẽga
sy no que la vista çiega
sy por algo que comienga
a mirar lexos se aluenga
entonçes muy menos llega
y queda tan mal librada
dela su loca porfia
que despues en sy tornada
apenas puede ver nada
ny lo poco que antes via

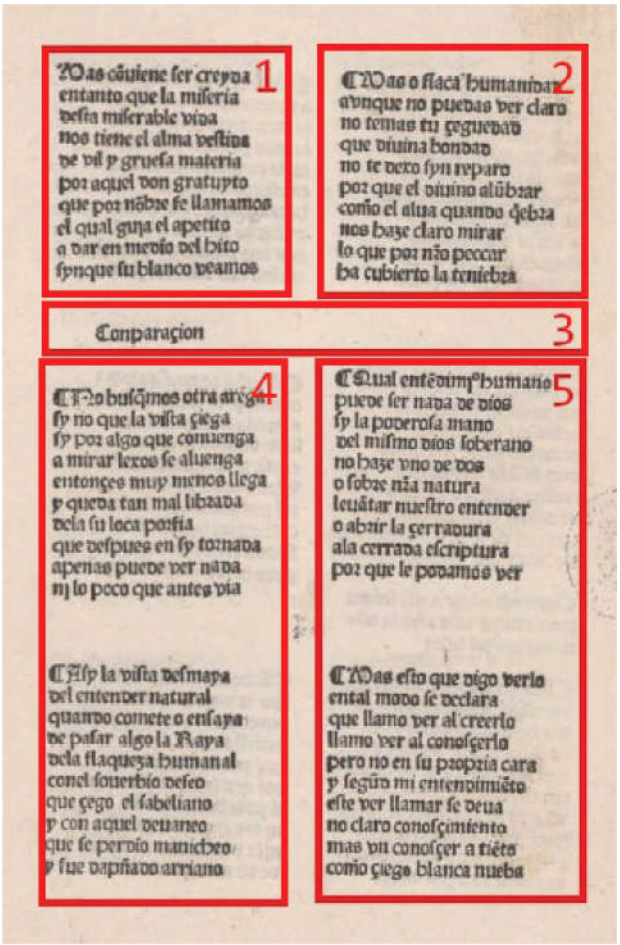
Qual entẽdimiẽto humano
puede ser nada de dios
sy la pobrosa mano
del mismo dios soberano
no haze vno de dos
o sobre nãa natura
leuãtar nuestro entender
o abrir la çerradura
ala cerrada escriptura
por que le podamos ver

Asy la vista desmaya
del entender natural
quando comete o ensaya
de pasar algo la Raya
dela flaqueza humanal
conel sobervio deseo
que çego el sabeliano
y con aquel detranco
que se perdio manicheo
y fue baptiçado arriano

Mas esto que digo verlo
ental modo se declara
que llamo ver al creerlo
llamo ver al conoscerlo
pero no en su propria cara
y segũ mi entendimiẽto
este ver llamar se deua
no claro conoscimiento
mas vn conoscer a tiẽto
como çiego blanca nueba

versos y estrofas de ambas columnas, con la disfunción que provoca cuando aparece en una de ellas una rúbrica de sección o de estrofa, puesto que genera un extenso espacio en blanco. Ante tal circunstancia, el algoritmo de segmentación podría actuar de manera errática e interpretar, así, que el título abarca ambas columnas, con lo que se separarían ambas estrofas de las anteriores y acabaría reordenando erróneamente el texto, en un orden de lectura como el que refleja la Figura 18.

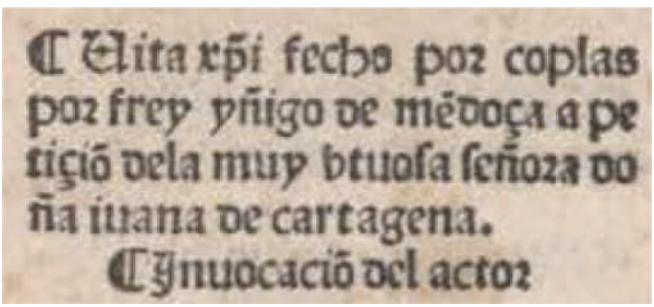
Figura 18. Segmentación y orden de lectura erróneo (Biblioteca Nacional de España, INC/2159, h. a5^{r} – 82IM)



En lo que concierne al reconocimiento de grafías, ya se ha discutido sobre la complejidad inherente de aplicarla a los tipos móviles góticos. Su elaboración manual y, por ende, la gran variedad existente, dificulta establecer un modelo de transcripción automática que se pueda aplicar de forma amplia a un gran número de obras de diferentes talleres o, incluso, del mismo taller, en lo que se denomina un *modelo extendido*.

Dado que las técnicas de impresión fueron perfeccionándose a lo largo de los años y, por tanto, los productos que salían de las prensas estaban cada vez más elaborados, resultaría de mayor interés para evaluar el software escoger un incunable temprano. Tanto su puesta en página como la tipografía aumentarán la complejidad de los procesos de segmentación y reconocimiento de caracteres por los motivos anteriormente mencionados. En este sentido, por tanto, la edición más antigua conocida de las *Coplas de la vita Christi* de fray Íñigo de Mendoza (82IM) (Zamora, Antón de Centenera, 1482), habría sido nuestra primera opción, por ser la *princeps*, si no fuese porque la digitalización que suministra la BNE es de escasa calidad y condicionará de forma artificial las características del software de reconocimiento (Figur*a 19*), razón por la que, finalmente, se ha decidido utilizar la edición inmediatamente posterior (82*IM) (Zaragoza, Juan Planck y Pablo Hurus, *c.* 1482), cuya captura es de mayor resolución.

Figura 19. Detalle de la visualización del texto (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)



No obstante, aunque ambas ediciones mantienen una puesta en página similar, varía la amplitud o extensión de líneas de su espacio interestrófico, tal como se evidencia al contrastar ambas ediciones plana a plana, lo que viene provocado por dos circunstancias que las diferencian: el formato y la tipografía. Por un lado, se pasa de una edición en 4º (82IM) a otra en folio (82*IM), mientras que, por el otro, también se aumenta el tamaño de la

tipografía, que pasa de una 75G a una 104G. Aunque podríamos pensar que, si el software consigue segmentar correctamente la edición 82*IM, lo haría también con la *editio princeps*, no sería así, necesariamente, por las diferentes circunstancias que concurren en ambos incunables, tanto de espacios interestróficos descompensados, como de falta de nitidez de sus digitalizaciones. Este hecho es extrapolable en mayor medida al resto de ediciones, con las que ni siquiera se comparte la misma distribución estrófica por plana, sino que el único punto en común es la composición a dos columnas. Esta circunstancia prioriza el comportamiento del software en esta fase de segmentación como uno de los criterios fundamentales para valorar en su elección, con el fin de evitar, en la medida de lo posible, un trabajo añadido a la transcripción, aunque no será el único, como veremos en el siguiente epígrafe.

5.2.3. Criterios de selección

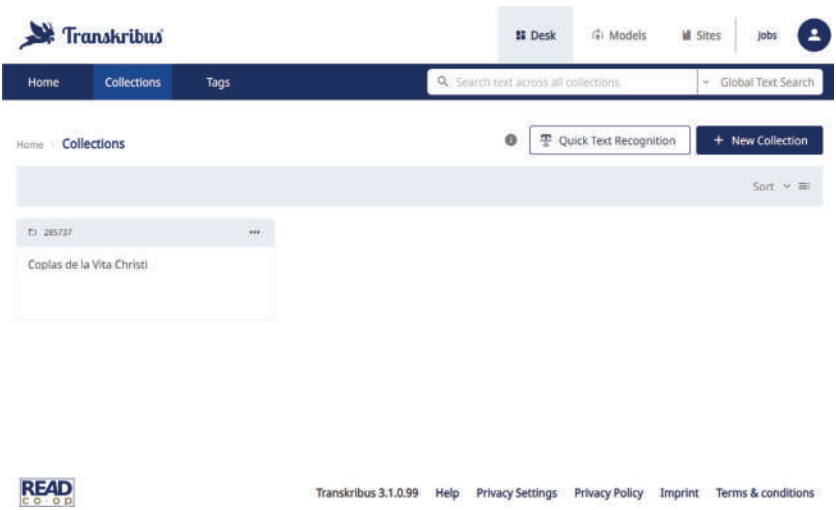
El análisis comparativo de softwares pivotará sobre cuatro características que nos definirán su madurez: el mecanismo de instalación, su usabilidad, el proceso de segmentación de las páginas con posibilidad de crear modelos personalizados para mejorarlo y, finalmente, el resultado de la transcripción con un modelo entrenado *ad hoc*.

Una instalación sencilla facilita en gran medida la adopción de un software, más aún si su público objetivo son personas que no son profesionales informáticos. Requerir la intervención de un experto para poner en marcha una aplicación ahuyentará a los usuarios más profanos en la materia y a los equipos de trabajo que aún no han dado el paso hacia la multidisciplinariedad.

El equipo de *Transkribus* debe tener clara esta premisa, ya que no requiere proceso de instalación, a diferencia de su antigua versión de escritorio, y en estos momentos únicamente se ofrece en versión web, al menos para el usuario común (Figura 20). Han seguido los pasos de otras grandes empresas, como es el caso de Microsoft, que ya dispone desde hace tiempo de una versión en línea de su famosa suite de ofimática *Office*, la *365*, accesible desde cualquier dispositivo con navegador. Las ventajas son evidentes, dado que se puede empezar a utilizar desde el primer momento sin pasar por un tedioso proceso de instalación, puesto que únicamente debemos disponer de un navegador reciente que soporte los últimos estándares web, lo que se traduce en que se podrá acceder tanto desde un teléfono inteligente, como desde un ordenador de sobremesa. Asimismo, todo nuestro trabajo quedará guardado en sus servidores, de manera que la migración de un dispositivo a otro es totalmente transparente y automática. Se puede empezar a trabajar en un proyecto en casa con un ordenador de sobremesa,

continuar en la cafetería con una tableta digital y finalizar la transcripción en el despacho con un portátil. Es evidente que las aplicaciones en la nube simplifican el desarrollo de proyectos al no estar atadas a una máquina en concreto, aunque también tienen sus inconvenientes, como son la necesidad de tener una conexión a Internet estable y rápida o, sobre todo, la dependencia absoluta de la empresa proveedora. Si sus servidores se apagan, nos quedamos sin datos; si deja de prestar el servicio, nos quedamos sin datos; si sufren un ciberataque, todos nuestros datos se venderán en Internet o, incluso, se ofrecerán de forma gratuita.

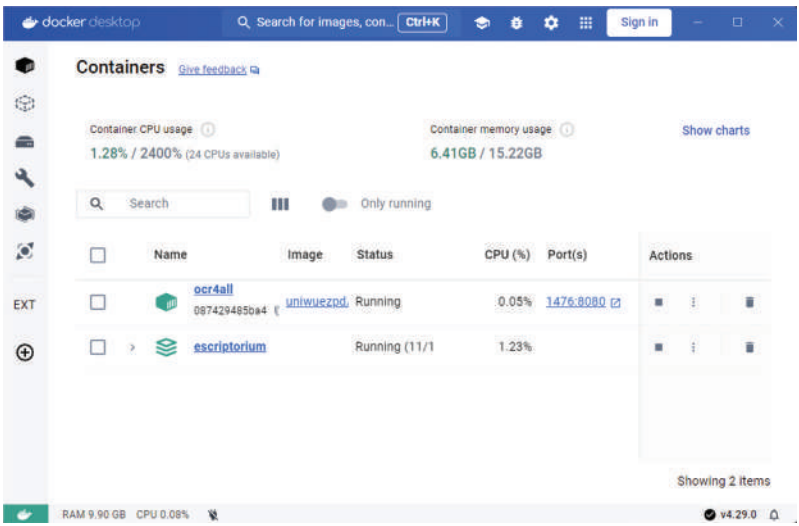
Figura 20. Interfaz de *Transkribus*



Por su parte, *OCR4all* tiene un proceso de instalación que, aunque no resulta complejo, sí que requiere un conocimiento técnico más allá del que suele ser habitual para una aplicación común y es que, en verdad, no es un software de escritorio al uso. Su funcionamiento, al igual que *Transkribus*, se basa en una interfaz web y, por tanto, se utiliza mediante un navegador. La complejidad de instalación se deriva de este formato de uso, al requerir aplicaciones adicionales de apoyo más allá de la propia aplicación de transcripción, como son un servicio web y una base de datos. A cambio, una vez instalada en un ordenador, se puede acceder a ella desde cualquier otro dispositivo que esté conectado a él siempre que permanezca encendido, por lo que su hábitat natural es un servidor.

El equipo de desarrollo ha simplificado esta complejidad inicial ofreciendo dos modalidades de instalación, una de ellas utiliza como base un sistema operativo *Linux* y la otra, mucho más rápida de llevar a cabo, se basa en utilizar un contenedor *Docker*. Aunque *Linux* es muy utilizado en entornos de servidor y es la base del sistema *Android* de los teléfonos móviles, su uso como entorno de escritorio está en torno al 4% a nivel mundial, en un mercado dominado de forma abrumadora por *Microsoft Windows* y seguido de lejos por *MacOS*[248]. Por ello, no es común que el usuario medio disponga de este sistema operativo para instarlo. La alternativa es el uso de un contenedor *Docker* (Figura 21), que precisará de conocimientos sobre este tipo de tecnología. *Docker* es un software que nos permite la ejecución de sistemas informáticos completos, a modo de servidor, independientemente del sistema operativo sobre el que estemos trabajando. Aunque su nivel de complejidad técnica no es elevado, no es habitual emplear este tipo de entorno fuera de los equipos de desarrollo y prueba de aplicaciones; sin embargo, simplifica la instalación de este tipo de software que requiere servicios auxiliares para su funcionamiento, como, por ejemplo, bases de datos.

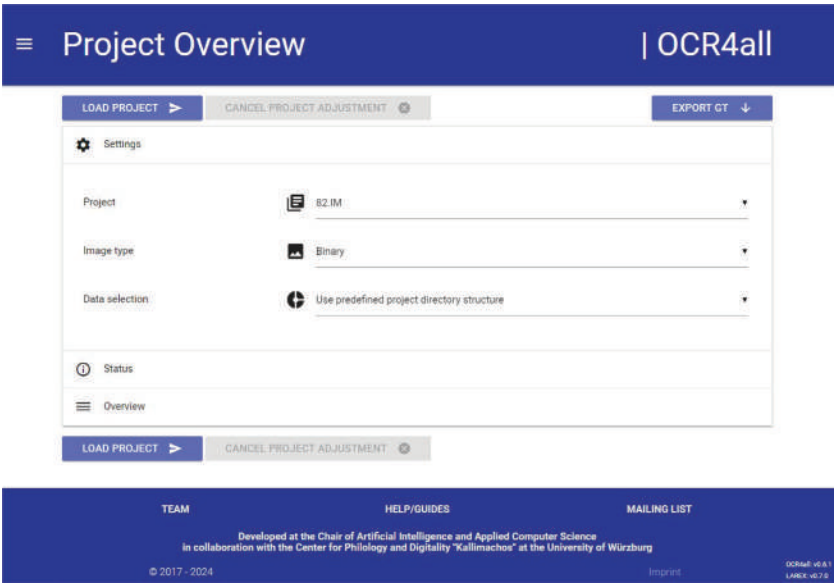
Figura 21. *Docker* ejecutando *OCR4all* y *eScriptorium*



248 Los datos están extraídos de la web de referencia *statcounter* y se pueden consultar en https://gs.statcounter.com/os-market-share/desktop/worldwide [consulta: 06/05/2024].

Entre las dos alternativas, se ha elegido la instalación mediante *Docker* por ser un proceso más rápido y sencillo que hacer una instalación completa en *Linux*. Para ello, se ha seguido la guía disponible en la página web del software para *Microsoft Windows*[249]. De forma resumida, los pasos han consistido en la creación de la estructura de directorios en el disco local, la descarga de la imagen de *Docker* que contiene todo el software necesario y, finalmente, la ejecución de la aplicación con los parámetros adecuados para que encuentre los archivos con los modelos y las imágenes a transcribir. Tras seguir esta secuencia, se pondrá en marcha el software que nos permite acceder a la interfaz de *OCR4all* a través del navegador y que nos permitirá crear nuestro primer proyecto de transcripción, tal como se muestra en la Figura 22.

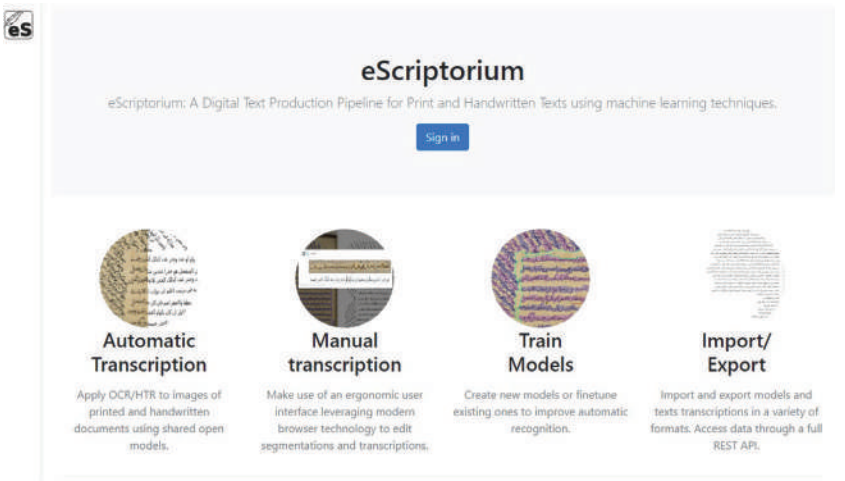
Figura 22. Interfaz de *OCR4all*



249 Existe una guía de instalación disponible en https://www.ocr4all.org/guide/setup-guide/quickstart [consulta: 06/05/2024].

Esta misma filosofía de desarrollo basada en un servidor web al que se accede a través de un navegador también ha sido seguida por *eScriptorium*. Sus ventajas son evidentes, puesto que se logra una total independencia del sistema operativo del usuario, además de que únicamente requiere un navegador, una aplicación que tiene en la actualidad cualquier dispositivo conectado a Internet, aunque, eso sí, se sacrifica la facilidad de instalación de las tradicionales aplicaciones de escritorio; de hecho, el proceso de instalación es casi idéntico en todos estos casos. También se ofrecen dos versiones: una instalación con todos los componentes de forma independiente en *Linux* y otra ya integrada, lista para ejecutar en formato *Docker*. Al igual que en *OCR4all*, se ha elegido la versión en *Docker* y, después de seguir los pasos de instalación que se ofrecen en la página web del software, también muy similares, se ha puesto en marcha la aplicación y se ha obtenido acceso a ella a través del navegador, cuya interfaz se muestra en la Figura 23.

Figura 23. Interfaz de *eScriptorium*



En el apartado de usabilidad, *Transkribus* tiene una interfaz minimalista muy intuitiva. La localización de las distintas funcionalidades es clara y se acompaña de una secuencia lógica de pasos. El usuario en ningún momento se siente perdido o abrumado entre múltiples alternativas, al mostrar únicamente en pantalla las opciones adaptadas al paso en el que

nos encontremos. Su funcionamiento se basa en lo que denomina *colecciones*, que no es más que una forma de agrupar documentos, por ejemplo, de temática similar o las distintas ediciones de la misma obra. Este concepto se asemeja al de una carpeta en nuestro disco duro local. Dentro de una colección colocaremos las imágenes de los documentos que vayamos a transcribir, con la posibilidad de subir los archivos en formato JPG, PNG o PDF mediante la acción de arrastrar y soltar. Para llevar a cabo la transcripción, únicamente habrá que seleccionar las páginas deseadas, pulsar el botón correspondiente y, de forma automática y sin intervención, se efectuarán todos los pasos. Una vez finalizado el proceso, nos mostrará la segmentación y la transcripción en paralelo en la misma página.

Esta interfaz nos permite la interacción entre ambos procesos en tiempo real, de manera que, si seleccionamos una línea en la imagen segmentada, esta aparecerá resaltada en el texto transcrito. Asimismo, permite corregir los errores que se hayan producido en cualquiera de ellas, tanto a nivel de detección de zonas textuales como de confusión de grafías, todo ello de forma sencilla e intuitiva con el uso del teclado y el ratón. El cuidado diseño visual que ofrece *Transkribus* tanto de distribución espacial como de iconografía da como resultado un producto profesional que facilita el empleo a los usuarios desde el primer momento en el que acceden. La curva de aprendizaje es mínima y es posible empezar a transcribir documentos después de unos pocos clics de ratón.

Por su parte, en la interfaz de *OCR4all* se ha prestado más atención a las posibilidades de personalización que a su atractivo visual. Aunque también sigue una estructura lógica para efectuar los pasos de la transcripción, su uso es menos intuitivo. Su funcionamiento se basa en proyectos, representados como una carpeta en nuestro dispositivo de almacenamiento con las imágenes del volumen con el que vamos a trabajar. No obstante, la transcripción deja de ser un proceso automático y, aunque también permite ejecutar todos los pasos de forma consecutiva, la selección de las opciones de cada uno de ellos se presenta con una interfaz parca y poco atractiva. Para efectuar la segmentación utiliza la aplicación *Kraken*, y su corrección se lleva a cabo de forma independiente con otra herramienta llamada *Larex* que, aunque es potente por las múltiples opciones que tiene, llega a abrumar y conlleva que el usuario se sienta perdido e incluso tenga que recurrir a un tutorial de uso. Por su parte, para la transcripción utiliza otra aplicación llamada *Calamari* con varios modelos ya instalados por defecto. Pese a que ninguno de ellos está realmente adaptado a la letra gótica, existe el de letra *Fraktur*, dado su uso más habitual en los impresos alemanes. El proceso de corrección de la transcripción presenta una interfaz también mucho más

árida respecto a la de *Transkribus*. Utiliza la misma herramienta que para la corrección de la segmentación, *Larex*, que muestra, línea a línea, la imagen original con la transcripción en la parte inferior.

La interfaz de *eScriptorium* se sitúa a medio camino entre la escasez de diseño empleado en *OCR4all* y la perfecta puesta en página de *Transkribus*. También funciona por proyectos, pero su concepto de uso se asemeja a las colecciones de *Transkribus*: una forma de reunir un grupo de digitalizaciones. Aunque carece de iconografía atractiva con tendencia a la simplificación, la distribución de las opciones a lo largo de la página favorece la usabilidad y permite que su manejo sea intuitivo. Para llevar a cabo los procesos de segmentación y transcripción utiliza *Kraken*, cuyos desarrolladores están integrados en el mismo proyecto. Aunque presenta la opción de efectuar el proceso de transcripción de manera automática, brinda la posibilidad de ajustar cada uno de los pasos (binarización, segmentación, transcripción), al igual que *OCR4all*. Sin embargo, los parecidos terminan aquí. Su interfaz está mucho más cuidada, con un manejo más intuitivo y que permite su uso sin apenas esfuerzo. Muestra de ello es la página diseñada para corregir tanto la segmentación como la transcripción, que presenta de forma paralela la imagen segmentada y el texto interpretado, al igual que hace *Transkribus*.

Pese a las diferencias a nivel visual, las tres aplicaciones comparten una de las características fundamentales: la posibilidad de entrenar un modelo de transcripción para mejorar el reconocimiento de grafías. No ocurre lo mismo con la funcionalidad de creación de un modelo de segmentación. *OCR4all* carece de esta posibilidad a través de la interfaz, lo que limita su uso en el caso de que el material con el que se trabaje tenga una puesta en página problemática, tal como sucede con las ediciones incunables de las *Coplas de la vita Christi*.

5.2.4. Entrenamiento y prueba de un modelo individual

Dadas las especiales características de la letra gótica española y para que los tres estuviesen en igualdad de condiciones, se entrenó un modelo individual con cada uno de ellos con la edición elegida, la zaragozana de Planck y Hurus (82*IM). Para ello, se generó de manera automática la transcripción de un determinado número de páginas que luego se corrigieron manualmente para formar lo que se denomina *Ground Truth*, esto es, el conjunto formado por las imágenes y su correspondiente transcripción corregida que se tomarán como base para entrenar un modelo. El *Ground Truth* que se generó estaba formado por 25 planas de la edición cuyo contenido

textual sumaba más de 5000 palabras, tal como se recomienda en la página de *Transkribus* para modelos de reconocimiento de material impreso[250].

En el primer paso a comparar, la segmentación, ya surgieron las primeras diferencias. Así como *Transkribus* detectó perfectamente las dos columnas de todas las páginas con el orden de lectura correcto de los versos, en el caso de *OCR4all* y *eScriptorium* el resultado fue, realmente, muy desacertado y, por tanto, inaceptable, con un número notable de errores en la mayor parte de páginas, sin duda derivados de que ambos utilizan el mismo software para llevar a cabo la segmentación, *Kraken*. En diversas ocasiones, el algoritmo aglutinó en un mismo campo de texto la primera estrofa de cada columna, lo que generó un resultado que alteraba la ordenación textual, no solo en lo que respecta a la secuencia de estrofas, sino también en la propia combinación de versos, procedentes de una y otra, ya que interpretaba el texto a línea tirada. Esta segmentación se tuvo que generar de nuevo manualmente, aunque conviene apuntar que esta corrección implicó un proceso lento al tener que eliminar la segmentación errónea, para crear la nueva y establecer, así, el orden adecuado de lectura de las diferentes zonas textuales.

Como ya se había anticipado, las especiales características de la puesta en página de la poesía medieval podrían dificultar la correcta detección de las zonas de texto, como así ocurrió. La ausencia de una alineación en el margen derecho del texto, así como la escasa separación entre las columnas en algunos casos, provocó que el algoritmo combinase dos zonas que pertenecían a columnas distintas en una sola.

Con el fin de salvar esta situación, se decidió entrenar un modelo de segmentación y, dado que *OCR4all* no tiene esta opción, se llevó a cabo con *eScriptorium*. Para generar este nuevo modelo de segmentación adaptado se emplearon las mismas 25 páginas utilizadas para el modelo de transcripción, pero en este caso, segmentando manualmente las columnas y estableciendo el orden de lectura correcto. Con estos datos, se obtuvo

250 En el manual de *Transkribus* se detallan los requisitos para generar modelos para material manuscrito e impreso: "Depending on the type of material and the number of hands, between 5,000 and 15,000 words (around 25-75 pages) of transcribed material are required to start. In general, the neural networks of the Text Recognition engine learn quickly: the more training data they have, the better the results will be. If you are working on printed material, 5,000 words should be sufficient to achieve a good Character Error Rate". Para más detalles consultar https://help.transkribus.org/data-preparation [consulta: 08/05/2024].

un modelo con una precisión que no llegaba al 50%, lo que teóricamente daría como resultado que la mitad de páginas sobre las que se aplicase tuviesen una segmentación incorrecta. Este comportamiento se verificó empíricamente segmentando el resto de páginas de la edición. Aunque algunas de ellas estaban bien segmentadas, aproximadamente la mitad requerían de una intervención manual para corregir la unión de zonas textuales de columnas distintas.

Previamente a la generación del modelo individual de transcripción, se efectuó la transcripción utilizando los modelos suministrados por cada software, a fin de obtener datos para verificar si se obtenía una mejora con el entrenamiento. En el caso de *Transkribus*, los modelos seleccionados fueron *Print M1*[251], creado por el mismo equipo que ha desarrollado el software a partir de material impreso en varias tipografías, incluyendo las góticas, y en diversos idiomas, entre los que se encuentra el español[252]; y *SpanishGothic_XV-XVI_extended_v1.2*, generado por un equipo internacional de investigadores, coordinado por Bazzaco, utilizando impresos en prosa española de finales del XV y del XVI[253]. En *OCR4all*, se seleccionó el modelo *deep_fraktur_hist*, entrenado en textos históricos con tipografía *Fraktur*, ante la ausencia de uno centrado en gótica incunable. Por su parte, dada la carencia de modelos instalados por defecto de *eScriptorium*, se buscaron por Internet aquellos compatibles con *Kraken*, entre los que se seleccionó uno supuestamente adaptado a impresos en gótica

251 Según la descripción del modelo que figura en la web de *Transkribus*: "Extended multilanguage Transkribus print model, including antiqua and blackletter prints, typewriter, computer print outs and decorative fonts Includes more languages than print 0.3. The CER in M1 is higher than in 0.3 which is due to a more varied validation set. For languages that were already included in 0.3 the new M1 usually performs equally well or slightly better than 0.3. Curated by the Transkribus team, this model is occasionally updated with community data for continuous improvement".

252 O, siendo estrictos, el castellano medieval y sus estados lingüísticos del Siglo de Oro.

253 Según la descripción del modelo que figura en la web de *Transkribus*: "Model created by: Stefano Bazzaco (coord.) Nuria Aranda García Ángela Torralba Ruberte Pedro Monteiro Giada Blasut Federica Zoppi Ana-Milagros Jiménez José Manuel Fradejas Eduardo Camero Santos Laura Lecina Nogués Almudena Izquierdo Andreu documents type: Printed typeface: Spanish Gothic (XV-XVIC) based on: Historia de la linda Magalona, Anónimo (Sevilla, Jacobo Cromberger, 1519) Historia de la reina Sebilla, Anónimo (Burgos, Felipe de Junta, 1551) Historia del rey Canamor, Anónimo (Valencia, Jorge Costilla, 1527) Libro del conde Partinuplés (Sevilla, Jacobo Cromberger, 1519) Libro del conde Partinuplés (Burgos, Herederos de Juan de Junta, 1558)

denominado *CATMuS Gothic Print*[254], disponible en el repositorio de modelos *Zenodo.*

La transcripción obtenida con los dos modelos de *Transkribus*, *Spanish-Gothic_XV-XVI_extended_v1.2* y *Print M1* y con el de *OCR4all*, *deep_fraktur_hist*, fue similar en cuanto a número de fallos, con un porcentaje muy bajo. Por el contrario, la salida de *CATMuS Gothic Print* no fue satisfactoria, dado que comportaba un gran número de errores en comparación con los resultados de los otros modelos. En *eScriptorium*, además de una segmentación y orden de lectura de versos erróneos, la transcripción producía un resultado extraño ya desde la primera línea:

'Tit a xp̃i secho p̊r cepla' en lugar de 'Uita xp̃i fecho por coplas'.

El mejor resultado se consiguió con el modelo *Print M1* de *Transkribus*, aunque la diferencia fue mínima. Pese a que el número de errores es similar, no ocurre lo mismo con el tipo de edición obtenida. Mientras que en el caso de los modelos de *Transkribus* se obtiene una transcripción

---

Libro del conde Partinuplés (Burgos, Felipe de Junta, 1563) Doctrinal de los Caballeros, Alonso de Cartagena (Burgos, Fadrique Biel de Basilea, 1487) La Fiameta, Juan Boccaccio (Salamanca, [Impresor de la Gramática de Nebrija], 1497) Crónica del Rey Don Rodrigo (Crónica Sarracina), Pedro de Corral ([Sevilla], [Meinardo Ungut y Estanislao Polono], 1499) Silves de la Selva, Pedro de Luján (Sevilla, Dominico de Robertis, 1549) Leandro el Bel, Pedro de Luján (Toledo, Miguel Ferrer, 1563) Florando de Inglaterra, (Lisboa, Germán Gallarde, 1545) Lisuarte de Grecia, Feliciano de Silva (Sevilla, Jácome Cromberger, 1550) Lisuarte de Grecia, Juan Díaz (Sevilla, Jacobo y Juan Cromberger, 1526) Tragicomedia de Calisto y Melibea (Roma, Marcellus Silber, 1515) Retablo de la Vida de Cristo, Juan de Padilla (Sevilla, Juan Cromberger, 1510) Crónica del Cid (Fadrique Biel de Basilea, Burgos, 1512) Siete Partidas, Antonio Díaz de Montalvo (Polonio y Ungut, Sevilla, 1491) Siete Partidas, Francisco de Velasco (Gregorio de Gregoriis, Venecia, 1528) Valerio de las historias escolásticas y de España, Diego Rodríguez de Almela (Juan de Ayala, Toledo, 1541) transcription criteria: semi-diplomatic transcription, abbreviations: solved, accents: no."

254 *CATMuS Gothic Print*, desarrollado por Sonia Solfrini y Simon Gabay, "is a *Kraken* OCR model fine-tuned on *CATMuS Medieval* model with data produced by the *SETAF* project. *SETAF* data are prints in Gothic typefaces and in 16th century French language. Transcriptions follow graphematic principles and try to be as compatible as possible with guidelines previously published for French: no ligatures (except those that still exist), no allographetic variants, abbreviations are not resolved. The model is trained with NFD Unicode normalization: each diacritic (including superscripts) are transcribed as their own characters, separately from the 'main' character". Se puede obtener en la dirección https://zenodo.org/records/10599911 [consulta: 09/05/2024].

semipaleográfica, modernizando las grafías como la *s* larga, la *r rotunda*, el *et* tironiano y resolviendo las abreviaturas, el modelo *deep_fraktur_hist* de *OCR4all* lleva a cabo una transcripción paleográfica pura, manteniendo todas las grafías y abreviaturas del texto original. Asimismo, los modelos de *Transkribus* introducen espacios ausentes en el texto original a modo de corrección, como por ejemplo *enel* se transcribe como *en el*, así como la sustitución de la letra *j* por *i*, como sucede en los versos 3 y 4 de la hoja Aiij^{v} del impreso. Este comportamiento es provocado por las normas de transcripción que se han seguido para crear el *Ground Truth* y que limitan o, incluso, invalidan su utilización cuando estas normas cambian, como es el caso y como se verá más adelante.

Como primera prueba de concepto para comparar el comportamiento de estos modelos entrenados se utilizó la primera hoja del ejemplar que se conserva en la Real Biblioteca del Monasterio de El Escorial de 82*IM. Como ya se ha adelantado, el que mejor comportamiento tuvo fue *Print M1*, que consiguió interpretar correctamente el 99% de los caracteres, es decir, obtuvo un CER del 1%. *SpanishGothic_XV-XVI_extended_v1.2* y *deep_fraktur_list* consiguieron un CER de 1.6% y 1.4%, respectivamente. La siguiente tabla contiene cada una de las transcripciones con los distintos errores marcados en cada línea, así como la suma de ellos y el porcentaje que suponen respecto al total de caracteres transcritos.

Se puede observar que la transcripción del modelo *deep_fraktur_list* se ve penalizada al haber tenido un entrenamiento basado únicamente en textos alemanes, por lo que es incapaz de identificar correctamente la grafía *ç*, al no utilizarse en este idioma. El resto de errores son debidos, principalmente, a la escasa separación entre letras y a la confusión entre grafías similares (*u/n*, *r/x*). Cada modelo presenta problemas distintos motivados por el corpus de entrenamiento utilizado y las normas de edición seguidas. Un modelo individual debería conseguir mejores resultados y corregir estas pequeñas deficiencias provocadas por las tipografías utilizadas.

Siguiendo la recomendación de *Transkribus*, para generar este nuevo modelo adaptado, se creó el *Ground Truth* con las primeras 25 páginas de la edición, partiendo de la transcripción ofrecida por el modelo *Print M1* y aplicando al texto resultante las siguientes normas de edición:

- Se desarrollan las abreviaturas
- Se mantienen las grafías *u/v*, *i/j* y *c/ç*
- Se mantiene la *ñ*
- Se mantiene la separación o aglutinación de palabras, como en *dela*, *ala* y *alcançar la*

Tabla 3. Transcripción de la primera plana de 82*IM

Print M1		**SpanishGothic_ XV-XVI extended_v1.2**		**deep_fraktur_hist**	
Uita chrsti fecho por coplas	0	Uita christi fecho por coplas	0	Vita xp̃i ſecho por coplas	2
por frey yñigo de mendoça a	0	por frey yñigo de mendoça a	0	por frey yñigo de mendoca a	0
petiçio de la muy virtuosa se	1	petiçion de la muy virtuosa se	0	petició dela muy virtuoſa ſe	0
ñora doña juana de cartagena	0	ñora doña juana de caxtagena	1	ñora doña juana de cartagena	0
Inuocacion del actor	1	Inuocacion del ac_or-	3	Jnuocacion del ac_or	1
Aclara son dininal	1	aclara son diuinal	1	Aclara ſon dininal	1
la çerrada niebla obscura	0	la çerrada niebla obscuxa	1	la cerrada niebla obſcura	1
que en_el linaje humanal	1	que en_el linaje humanal	1	que enel linaje humanal	0
por la culpa paternal	0	por la culpa paternal	0	por la culpa paternal	0
desdel comienço nos dura	0	desdel comienço nos duxa	1	deſdel comienco nos dura	1
despierta la voluntad	0	despierta la voluntad	0	deſpierta la voluntad	0
endereça la memoria	0	endereça la memoria	0	endereca la memoria	1
por que syn contrariedad	0	porque syn contraxiedad	1	por que ſyn contrariedad	0
a tu alta magestad	0	a tu alta magestad	0	a tu alta mageſtad	0
se cante diuina gloria	0	se cante diuina gloria	0	ſe cante diuina gloria	0
Aquella grand compasyon	0	aquella grand compasyon	0	Aquella grand compaſyon	0
aquel amor entrañal	0	aquel amor entrañal	0	aquel amor entrañal	0
que por nuestra saluaçion	0	que por nuestra saluaçion	0	que por nueſtra ſaluacion	1
hizo sofrir tal passion	0	hizo sofrir tal passion	0	hiʒo ſofrir tal paſſion	0
atu fijo natural	0	a tu fijo natural-	1	atu fijo natural	0
aquella bondad diuina	0	aquella bondad diuina	0	aquella bondad diuina	0

Tabla 3. Transcripción de la primera plana de 82*IM

Print M1		**SpanishGothic_ XV-XVI extended_v1.2**		**deep_fraktur_hist**	
que le forço a ser ombre	1	que le forço a ser ombre	1	quele forco a ſer ombre	1
enmiende lo que se inclina	0	enmiende lo que se inelina	1	enmiende lo que ſe inclina	0
en esta carne mesquina	1	en esta carne mesquina	1	eneſta carne meſquina	0
a offender el tu nombre	0	a offender, el tu nombre	1	a cffender el tu nombre	1
Srosigue-	2	brosigue	1	Proſigue	0
Los altos mereçimientos	0	Los altos mereçimientos	0	Los altos merecimientos	1
de aquella virgen y madre	0	de aquella virgen y madre	0	de aquella virgen y madre	0
y los asperos tormentos	0	y los asperos tormentos	0	y los aſperos tormentos	0
que sufren por ti contentos	0	que sufren por ti contentos	0	que ſufren por ti contentos	0
los que te tienen por padre	0	los que te tienen por padre	0	los que te rienen por padre	1
y lauitoria famosa	0	y lauitoria famosa	0	y lauitoria famoſa	0
de tus martires pasados	0	de tus martires pasados	0	de tus martires paſados	0
me alcançen que la prosa-	1	me alcançen que la prosa-	1	me alcangen que la proſa	1
de tu vida gloriosa	0	de tu vida gloriosa	0	de tu vida glorioſa	0
escriua en metros rimados	0	escriua en metros ximados	0	eſcriua en metros rimados	0
Despide las musas poeticas	0	Despide las musas poeticas	0	Deſpide las muſas poeticas	0
e inuoca las cristianas	0	e inuoca las cristianas	0	e inuoca las criſtianas	0
Dexemos las poesias	0	Dexemos las poesias	0	Dexemos las poeſias	0
y sus musas inuocadas	0	y sus musas inuocadas	0	y ſus muſas inuocadas	0
por que tales ninirias	1	porque tales ninixias	2	por que tales nimrias	1

Tabla 3. Transcripción de la primera plana de 82*IM

Print M1		**SpanishGothic_ XV-XVI extended_v1.2**		**deep_fraktur_hist**	
por humanas fantasias	0	por humanas fantasias	0	por humanas fant_aſïas	1
son cierto temorizadas	0	son cierto temorizadas	0	ſon cierto temoriʒadas	0
y veniendo a_la verdad	1	y veniendo a_la verdad	1	y veniendo ala verdad	0
de quien puede dar ayuda	0	de quien puede dar ayuda	0	de quien puede dar ayuda	0
a_la sola trinidad	1	a_la sola trinidad	1	ala ſola trinidad	0
que mana siempre bondad	0	que mana siempre bondad	0	que mana ſiempre bondad	0
gela pidamos sin duda	0	gela pidamos sin duda	0	gela pidamos ſin duda	0
Prosigue	0	Prosigue	0	Proſigue	0
Non digo que los poetas	0	Non digo que los poetas	0	Non digo que los poetas	0
los presentes y passados	0	los presentes y passados	0	los preſentes y paſſados	0
non fagan obras perfectas	0	non fagan obras perfecas	0	non fagan obras perfectas	0
graçiosas y bien discretas	0	graçiosas y bien discretas	0	graçioſas y bien diſcretas	1
en sus renglones trobados	0	en sus xenglones trobados	1	en ſus rẽglones trobados	0
mas affirmo ser herror	0	mas affirmo ser herror	0	mas affirmo ſer herror	0
perdonen sy bien non fable	1	perdonen sy bien non fablo-	1	perdonen ſy bien non fablo	0
en su obra el trobador	0	en su obra el trobador	0	en ſu obra el trobador	0
inuocar al dios de amor	0	inuocar al dios de amor-	1	inuocar al dios de amor	0
para seruiçio del diablo	0	para seruiçio del diablo	0	para ſeruiçio del diablo	1
Prosigue y prueua-	1	Prosigue y prueua	0	Proſigue y prueua	0
con sant jheronimo	0	con sant jheronimo	0	con ſant jheronimo	0

Tabla 3. Transcripción de la primera plana de 82*IM

Print M1		**SpanishGothic_ XV-XVI extended_v1.2**		**deep_fraktur_hist**	
Sant jheronimo acusado	0	Sant jheronimo acusado	0	Sant jheronimo acuſado	0
por que en çiceron leya	0	porque en çiceron leya	0	por que en c̄iceron leya	1
en spiritu arebatado	0	en spiritu ax̄ebatado	1	en ſpiritu arebatado	0
fue duramente açotado	0	fue duxamente açotado	0	fue duramente ac̄otado	1
presente dios que le dezia	0	presente dios que le dezia	0	preſente dios que le deʒia	0
si piensas que eres p̄hristiano	1	si piensas que ex̄es christiano	1	ſi pienſas que eres x p̄iano	1
segund la forma deuida	0	segund la forma deuida	0	ſegund la forma deuida	0
es vn pensamiento vano	0	es vn pensamiento vano	0	es vn penſamiento vano	0
que ex̄es çiceroniano	1	que ex̄es çiceroniano	1	que eres c̄iceroniano	1
pues es çiceron tu vida	0	pues es çiceron tu vida	0	pues es c̄iceron tu vida	1
1589 cars. – CER 1%	16	1591 cars.– 1.6%	26	1575 cars. – 1.4%	22

- Se sustituyen las siguientes grafías:
  - 7 *(et tironiano)* por *e* (o *et* en expresiones y citas latinas)
  - ꝛ *(r rotunda)* por *r*
  - ſ *(s* larga) por *s*
  - ȝ *(ezh)* por *z*

A partir del *Ground Truth* en los tres softwares, se entrenó un modelo individual de transcripción en cada uno de ellos al que se llamó *Spanish Gothic Poetic Incunabula*. La creación de este modelo es distinta en cada software. En *Transkribus*, hay que seleccionar las páginas que forman el *Ground Truth* y la interfaz nos guiará de forma visual por los siguientes pasos. Primero hay que establecer el porcentaje de páginas que se utilizarán como verificación, que suele ser un 10% del total, y que servirán para aplicar el modelo sobre ellas y calcular el CER. En el siguiente paso hay que seleccionar, si se desea, un modelo preexistente del que partir y, finalmente, en el último paso se deben rellenar los datos descriptivos, como son el nombre del modelo, los siglos que abarca y si es para impresos o manuscritos. La generación del modelo en *OCR4all*, aunque no sea visual y guiada como en *Transkribus*, también es sencilla, ya que la aplicación suministra por defecto unos valores de configuración estándar y únicamente hay que seleccionar un modelo preexistente en el caso que se quiera esta opción. En el caso de *eScriptorium* es igual de sencillo, puesto que hay que seleccionar las páginas que se utilizarán para el entrenamiento, elegir un modelo ya existente, si es el caso, y darle un nombre. Aunque a la hora de generar un nuevo modelo todos los softwares dan la posibilidad de añadir el entrenamiento a uno ya existente para tomarlo como base, con el fin de evitar contaminaciones de normas de edición distintas, se decidió prescindir de esta opción y que el software generase un modelo desde cero en todos los casos.

Con el entrenamiento efectuado en *Transkribus* se obtuvo un modelo con un CER de 0.70%. Este resultado indica que está previsto que se produzca un error de transcripción de una grafía cada 143 correctas, aproximadamente. *OCR4all*, a diferencia de *Transkribus*, no genera un único modelo, sino varios de forma paralela, por lo que se decidió obtener cuatro, de manera que dispusiésemos de una muestra amplia aprovechando las posibilidades ofrecidas por este software. Sus CER variaron entre el 0.80% y el 0.46%, unas cifras que se asemejan a las obtenidas por *Transkribus*. Por último, el modelo generado por *eScriptorium* consiguió un CER de 31%, lo que contrasta con los otros modelos y lo hace inusable, al menos para la transcripción automática destinada a una edición paleográfica, crítica o actualizada, ya que sería necesario corregir un tercio del texto que se obtuviese. En la siguiente tabla se muestran de forma conjunta estos resultados:

Tabla 4. Porcentaje de errores de los modelos de transcripción

CER modelo *Transkribus*	**CER mod. 1 *OCR4all***	**CER mod. 2 *OCR4all***	**CER mod. 3 *OCR4all***	**CER mod. 4 *OCR4all***	**CER modelo *eScriptorium***
0.70%	0.80%	0.70%	0.55%	0.46%	31%

Tales resultados evidencian la similitud entre el motor de transcripción de *Transkribus* y el de *OCR4all* (*Calamari*), cuando se entrena un modelo individual con los mismos datos. Los porcentajes de errores son similares: ambos consiguen una transcripción casi perfecta y demuestran la mejora que se consigue con la generación de un modelo individual especialmente adaptado a los textos que se vayan a transcribir. Sin embargo, no ocurre lo mismo con la segmentación: la aplicación *Kraken*, utilizada tanto por *OCR4all* como por *eScriptorium*, no consigue aislar correctamente las zonas de texto y, por consiguiente, no proporciona una ordenación de los versos al final del proceso. Esta circunstancia era previsible, ante la necesidad de entrenar un modelo exclusivamente atendiendo a la segmentación por las características diferenciales de la puesta en página de la poesía incunable. No obstante, pese a llevar a cabo esta acción, tampoco se consiguió un comportamiento óptimo. El resultado obtenido, con la mitad de páginas erróneas, obligaba a un trabajo adicional que ralentizaba todo el proceso. Aunque es posible que, con un aumento del número de muestras utilizadas para el entrenamiento, se obtenga un mejor resultado, se considera que con la utilización de una quinta parte de hojas de la edición se debería de haber conseguido un resultado mejor. *Transkribus*, por el contrario, generó una segmentación perfecta en todas las hojas y, así, no fue necesario retocar ninguna zona textual. Por tanto, aunque *OCR4all* y *Transkribus* generaron modelos de transcripción similares en comportamiento y con excelentes resultados, el errático comportamiento de la segmentación de *OCR4all* inclinó la balanza hacia el uso de *Transkribus* como software para continuar con la investigación.

En definitiva, queda demostrado que el entrenamiento de una red neuronal con determinadas normas de edición a partir de la tipografía concreta de un testimonio de un incunable poético ofrece transcripciones con un porcentaje de error prácticamente nulo. Ahora bien, este comportamiento solo está asegurado si este modelo se aplica a ediciones con la tipografía con la que fue entrenado. Para vencer esta limitación, hay que seguir moldeando la red neuronal suministrándole más muestras con el fin de aumentar su conocimiento tipográfico. El objetivo último, la generación de un modelo extendido que transcriba incunables poéticos, únicamente

se puede alcanzar si el conjunto de tipos utilizados en el entrenamiento es suficientemente amplio como para resultar representativo de ese periodo.

El corpus de trabajo elegido, las *Coplas de la vita Christi* de fray Íñigo de Mendoza, ofrece un amplio muestrario de tipografías góticas del siglo xv, así como la mayor parte de rasgos materiales y problemas de transcripción automática que encontraremos en los incunables poéticos. El hecho de que las ocho ediciones incunables de esta obra provengan de distintos talleres impresores, permitirá la generación de un *Ground Truth*, o conjunto de entrenamiento, con el que obtener un modelo extendido con tasas de acierto cercanas también al 100%. No obstante, su verdadero valor no solo será su aplicación a la obra de referencia utilizada como corpus de entrenamiento, sino que, por el hecho de haberse utilizado un muestrario tipográfico que alcanza desde 1482 hasta 1499, esto es, una muestra diacrónica de prácticamente todo el arco de producción de impresos poéticos del siglo xv, se espera que su aplicación también se pueda extender a otros impresos poéticos incunables e, incluso, post-incunables, si no posteriores, con tasas de acierto similares.

5.3. La generación de un modelo extendido para la transcripción automática de incunables poéticos en tipografía gótica

Los modelos individuales que se generan al entrenar una red neuronal parten de una muestra específica, sin variación ni diversidad, por lo que su campo de aplicación acaba siendo muy limitado. En el caso que nos ocupa, si únicamente le proporcionamos ejemplos de una sola tipografía, será capaz de transcribir otras ediciones con esta misma letrería, pero tendrá dificultades para reconocer otras distintas, aunque para nosotros sean de apariencia similar. Es aquí donde entran en escena los modelos extendidos o, lo que es lo mismo, las redes neuronales entrenadas con una amplia variedad de ejemplos, con el objetivo de conseguir que aumente su campo de aplicación. En principio, si la red neuronal se ha sometido a un buen proceso de entrenamiento a partir de una amplia selección de tipografías, podrá reconocer formas más allá de aquellas para las que ha sido entrenada.

Un modelo extendido de transcripción se puede generar en un único paso, alimentando a la red neuronal con un *Ground Truth*, es decir, con un conjunto de imágenes con su correspondiente transcripción, que contenga una variedad tipográfica representativa; o bien de forma incremental, esto es, se va entrenando la red neuronal de forma secuencial con las diversas tipografías, una tras otra, obteniendo con cada paso un modelo entrenado más potente.

Aunque ambas filosofías de trabajo deberían obtener al final un resultado idéntico si se utilizan los mismos datos para ambas, la forma incremental de generación, aunque más laboriosa, nos mostrará en detalle su eficiencia con cada tipografía, esto es, su CER. Este valor, obtenido como porcentaje, indicará el número de errores que previsiblemente cometerá el modelo en cada transcripción. Dada la diacronía del corpus elegido, se ha decidido seguir este tipo de entrenamiento por la información que nos pueda aportar el comportamiento de la transcripción según el año de la tipografía que está tratando, un dato muy interesante por ser algunas de las tipografías más tempranas y arcaicas, tanto del propio siglo xv, como de cada uno de los talleres en cuestión. Tales dificultades añadidas ayudarán a la generación de un modelo extendido de mayor solidez y alcance, pues se amplía la casuística de aquellas que podrá encontrar en su aplicación a otras letrerías, incluidas las góticas del siglo xvi.

El modelo inicial de transcripción del que se parte, cuya elaboración se ha descrito en el epígrafe anterior, está entrenado únicamente con la tipografía GW-ID 1:104G utilizada en el testimonio 82*IM de las *Coplas de la vita Christi* de fray Íñigo de Mendoza. En concreto, se ha actuado a partir de la digitalización del ejemplar del Real Biblioteca del Monasterio de San

Lorenzo de El Escorial por su mejor estado de conservación y por la buena calidad de la digitalización. Ante la mala calidad de la reproducción de la *editio princeps* (82IM) y dada su similitud de puesta en página con 82*IM, se decide la elaboración del modelo individual a partir de esta última, para, después, entrenar un modelo extendido ya en el orden cronológico de los testimonios.

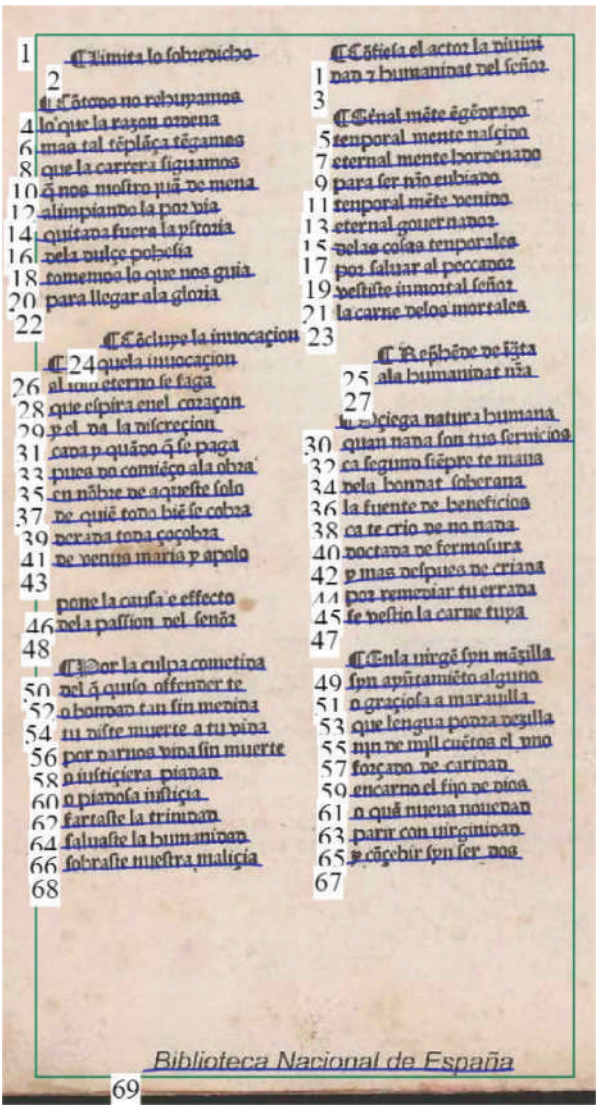
La digitalización del ejemplar único de 82IM, conservado en la BNE, presenta, en efecto, una baja calidad de captura, lo que, unido a su formato en 4º y al minúsculo tamaño de su tipografía, una 75G, acaba afectando irremediablemente a la nitidez de las grafías y, con ello, a su proceso de transcripción automática. Este es un problema de base importante, pero, en realidad, antes de reconocer su tipografía, ya aparecieron errores en el proceso de segmentación previo, que se ejecutó de manera errática en todas las páginas del impreso, agrupando las dos columnas en una sola y, por tanto, creando un orden de lectura incorrecto, tal como se observa en la Figura 24. Aunque no se sabe con seguridad a qué es debido este comportamiento, es muy posible que sea consecuencia de la marca de agua que la BNE ha incorporado a pie de página para indicar su procedencia, puesto que esta aparece centrada y confunde al algoritmo, lo que desemboca en la unión de ambas zonas textuales como si fuesen un texto a línea tirada.

La solución a este imprevisto obligó a una modificación manual de la segmentación de todas y cada una de las páginas, para crear dos zonas perfectamente diferenciadas que abarcasen ambas columnas, dejando fuera el texto contemporáneo superpuesto de la parte inferior. A todo ello se suma, además, que, a menudo, la caja tipográfica está inclinada por problemas de impresión, siendo este un incunable poético tan temprano, como se puede comprobar, por ejemplo, en las h. aj^{v} y d8^{v}, frente a las h. aj^{r} y d8^{r}, cuyo texto está alineado correctamente.

Como en este punto de la investigación ya se disponía de un modelo individual de transcripción llamado *Spanish Gothic Poetic Incunabula*, creado únicamente a partir del testimonio 82*IM, una vez solucionada la segmentación, se decidió ponerlo en práctica para ver su comportamiento con otras grafías para las que no había sido entrenado. Aunque 82IM y 82*IM mantienen una puesta en página similar, cada uno tiene su propia tipografía, de un tamaño muy diferente, en el primero de ellos identificada por el GW como la 1*:75G de Centenera y el segundo como la 1:104G de los Hurus, siendo ambas, como se puede comprobar, la primera de las utilizadas por ambos talleres.

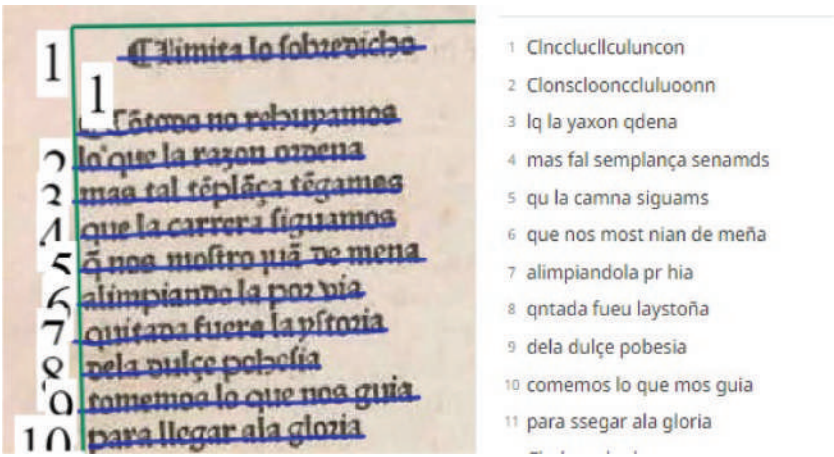
El resultado de la transcripción utilizando el modelo *Spanish Gothic Poetic Incunabula* sobre 82IM no fue bueno, ciertamente, a pesar de ser un modelo individual entrenado sobre una edición (82*IM) realizada a plana

Figura 24. Segmentación errónea (Biblioteca Nacional de España, INC/2159, h. aj^{v} – 82IM)



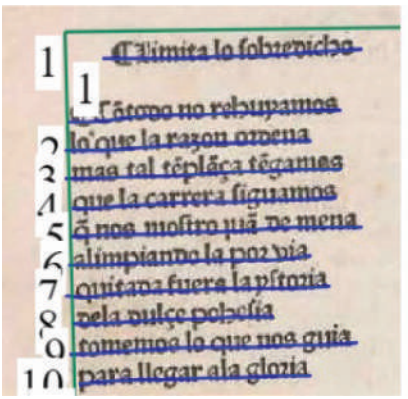
y renglón sobre la prínceps, lo que es buena muestra de que la aparente similitud morfológica de unas grafías para un ojo humano no es tal para la inteligencia artificial y, de esta manera, la diversidad tipográfica se reivindica como un criterio esencial para el desarrollo de un modelo extendido. Como se puede ver en el recorte de la Figura 25, la red neuronal es incapaz de identificar correctamente la mayor parte de las grafías, dando un resultado totalmente ininteligible.

Figura 25. Transcripción con *Spanish Gothic Poetic Incunabula* aj^{v} (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)



En vista del deficiente resultado obtenido con *Spanish Gothic Poetic Incunabula*, entrenado a partir de 82*IM, se optó por repetir el proceso inicial con 82IM, es decir, la fase de entrenamiento individualizado, a fin de obtener una transcripción que contribuyese con éxito a la creación de un modelo extendido. Para ello, se utilizó de nuevo el mismo modelo que se había utilizado para la transcripción inicial del 82*IM, el *Print M1* que *Transkribus* proporciona por defecto. El texto que se obtuvo (Figura 26), pese a no ser tampoco exacto, como ya se había previsto en el epígrafe anterior —y de ahí que se justifique acertada la decisión de usar 82*IM como base para la generación un modelo individual— fue más preciso que el anterior, por lo que se decidió tomarlo como punto de partida para transcribir las 25 primeras planas que formarían el *Ground Truth*.

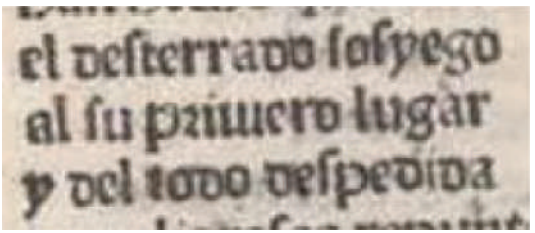
Figura 26. Transcripción con el modelo *Print M1* (Biblioteca Nacional de España, INC/2159, h. aj^{v} – 82IM)



1 Aunita la leburuca
2 Cawno un reruyanoos
3 lo que la razon ordena
4 mas tal templança tengamos
5 que la carrera siguamos
6 que nos mostro juan de mena
7 alimpiandola por via
8 quitada fuera la ystoria
9 de la dulçe pohesia
10 tomemos lo que nos guia
11 para llegar a la gloria

En la generación de este *Ground Truth* se observaron diversos errores de composición de esta edición que implicaron tomar una decisión en cuanto al texto resultante. El primero de ellos, y también el más curioso, se observa en la palabra *primero*, de la h. aiiij^{r}. El cajista se confundió y colocó el tipo de la letra *m* al revés, dando como resultado una especie de *w*, tal como se ve en centro de la Figura 27. Dado que no se trata de un error por confusión de tipos, puesto que el tipo era el de la *m*, en efecto, pero invertido, se decidió transcribir la palabra con su forma correcta.

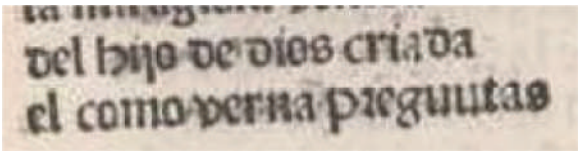
Figura 27. Tipo girado (Biblioteca Nacional de España, INC/2159, h. aiiij^{r} – 82IM)



Justo al final de esta estrofa, el cajista vuelve a cometer otra equivocación, aparentemente similar, aunque, en realidad, en este caso no es una inversión del tipo, sino una confusión, habitual, de hecho, entre la *u* y la *n*, un error por caja sucia, sin duda, al haber sido guardado en el cajetín de otro similar. Así se puede ver en la Figura 28, pero es una confusión que

se produce en varios puntos de esta edición, así como en otras ediciones. Como es lógico, porque buscamos una transcripción paleográfica, se optó por no corregir estos errores mecánicos, lo que, además, hubiese repercutido negativamente en el aprendizaje de la red neuronal, al proporcionarle la transcripción corregida de una grafía que no se identificaría con la que encontraría esta IA.

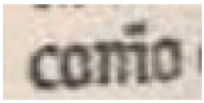
Figura 28. Confusión del tipo *u* por el tipo *n* (Biblioteca Nacional de España, INC/2159, h. aiiij^{r} – 82IM)



Una de las características que se repite a lo largo de toda la edición son las abreviaturas. Aunque las nasales son las comunes y compartidas por el resto de testimonios, incorpora otras menos habituales, como es el caso de la terminación *miento* por *j*^{o} (Figura 29), que se utiliza en varias ocasiones a lo largo del texto en palabras como *entendimiento* y *casamiento*. Respecto a las abreviaturas de las nasales y por su característica lingüística distintiva, cabe destacar su utilización de forma sistemática en la palabra *commo* (Figura 30).

Figura 29. Abreviatura en h. a5^{r} (Biblioteca Nacional de España, INC/2159, h. a5^{r} – 82IM)

Figura 30. Abreviatura de *commo* en h. a5^{v} (Biblioteca Nacional de España, INC/2159, h. a5^{v} – 82IM)



Asimismo, el texto del ejemplar ha sido alterado en diversos puntos, posiblemente con la intención de subsanar un error de impresión. En la mayor parte de casos, las grafías originales han sido rascadas y, en su lugar, aparece escrita a tinta una nueva grafía (Figura 31) o, incluso, varias (Figura 32). Como no lo podemos reconstruir a ciencia cierta, al ser el único ejemplar conservado, para su transcripción se ha partido del texto de la edición posterior, 82*IM, al ser una copia a plana y renglón de esta y, por tanto, es de esperar que con un texto idéntico cuando se trate de palabras completas que han sido manipuladas por un lector (*humildad/humanidad*), aunque en grafías concretas, como en el caso de *que*, sería muy posible que la corrección buscase revertir un error mecánico por confusión de tipos y es imposible saber cuál sería el empleado en la edición de Centenera:

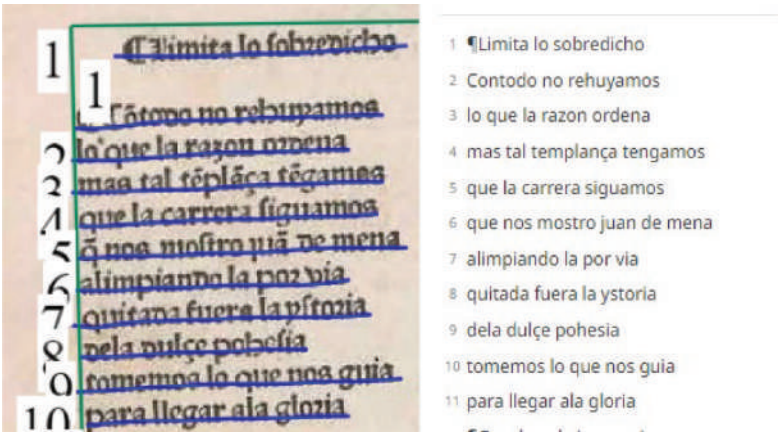
Figura 31. Rascado y corrección con la palabra *que* (Biblioteca Nacional de España, INC/2159, h. aiij^{r} – 82IM)

Figura 32. Rascado y corrección de la palabra *humildad* por *humanidad* (Biblioteca Nacional de España, INC/2159, h. aiij^{r} – 82IM)

A partir de las 25 planas transcritas, se generó un nuevo modelo, aunque partiendo del anterior, el *Spanish Gothic Poetic Incunabula* inicialmente creado, lo que en realidad conlleva ampliarlo para que sea capaz de reconocer una nueva tipografía y, por tanto, pase a ser un modelo extendido. Para conservar los modelos anteriores a efectos de comparación, se optó por crear versiones y llamarlo *Spanish Gothic Poetic Incunabula v.2*. En este caso, arrojó una tasa de error de un 2% en su aplicación a este ejemplar, algo previsible dada la mala calidad de su digitalización y el carácter temprano de la edición. No obstante, pese a este ligero incremento, el resultado mejora de forma considerable los modelos de los que se partía. En la Figura 33 se muestra, a modo de ejemplo, el mismo fragmento que

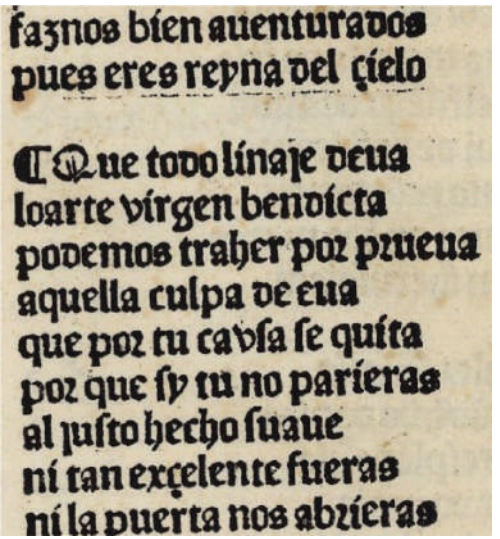
aparece en la Figura 25 y en la Figura 26, pero transcrito con este nuevo modelo. Si se comparan estos resultados, se evidencia claramente que, gracias al entrenamiento efectuado, ahora se obtiene una transcripción perfecta y adaptada a las normas de edición establecidas.

Figura 33. Resultado final con la versión 2 del modelo



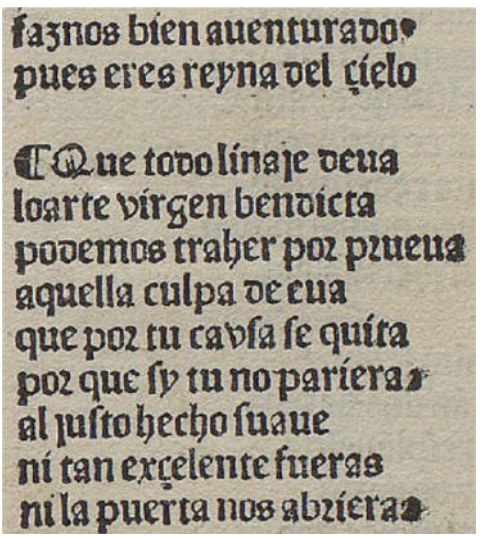
La siguiente edición que contiene las *Coplas de la vita Christi*, identificada por Dutton como 83*IM y que vio la luz poco después de 82*IM, también viene encabezada con esta obra, pero, a diferencia de los dos incunables poéticos previos, nos encontramos ya ante el primer cancionero impreso estrictamente castellano (Martos, 2018b, p. 528), salido también de las prensas zamoranas de Antonio de Centenera. De este impreso, afortunadamente, hay localizados tres ejemplares con un estado de conservación relativamente bueno, aunque el que se encuentra en la British Library es el mejor conservado y digitalizado, por lo que ha sido el utilizado para formar el *Ground Truth*. De este, sin embargo, hay que hacer notar que la letra *s* al final de algunos de los versos aparece bajo un borrón de tinta, que parece un problema de impresión original, resuelto en otros ejemplares, como el de la BNE, tal y como se hace evidente al comparar las palabras *auenturados*, *parieras* y *abrieras* (Figura 34) con la solución del ejemplar británico (Figura 35).

Figura 34. Texto no emborronado (Biblioteca Nacional de España, INC/897, h. a5^{r} – 83*IM)


faznos bien auenturados
pues eres reyna del çielo

¶Que todo línaje deua
loarte vírgen bendicta
podemos traher por prueua
aquella culpa de eua
que por tu cavſa ſe quita
por que ſy tu no parieras
al juſto hecho ſuaue
ní tan excelente fueras
ní la puerta nos abrieras

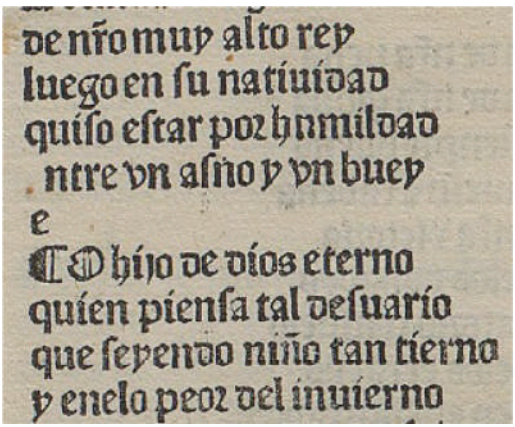
Figura 35. Texto emborronado (British Library, IB.52920, h. a5^{r} – 83*IM)


faznos bien auenturados
pues eres reyna del çielo

¶Que todo línaje deua
loarte vírgen bendicta
podemos traher por prueua
aquella culpa de eua
que por tu cavſa ſe quita
por que ſy tu no parieras
al juſto hecho ſuaue
ní tan excelente fueras
ní la puerta nos abrieras

Si bien es habitual que un ejemplar haya sufrido algún tipo de manipulación a manos de un antiguo poseedor o vendedor para modificar u ocultar determinados rasgos o características textuales, no lo son tanto los errores de impresión, como es el caso de la primera letra de un verso de la página a5^{v}, que, en lugar de aparecer en su renglón correspondiente, está justo en el inferior, en el espacio interestrófico (Figura 36).

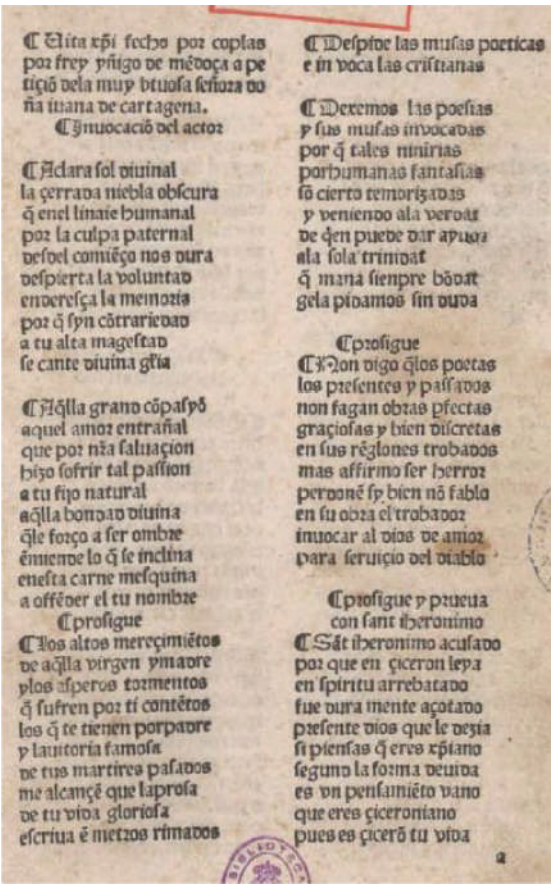
Figura 36. Tipo en el renglón inferior (British Library, IB.52920, h. a5^{v} – 83*IM)



Curiosamente, esta edición (83*IM), pese a haber salido del mismo taller zamorano, no comparte ya la impecable composición de la *editio princeps* (82IM), reproducida, asimismo, en la primera de Pablo Hurus (82*IM)[255], sino que, a partir de este momento, los cajistas no distribuyen las estrofas completas por columnas y están más preocupados en intentar incorporar el mayor texto posible en una hoja y, así, aprovechar al máximo el papel. Esta circunstancia se observa claramente ya en la primera hoja de 82IM (Figura 37) al compararla con la correspondiente de 83*IM (Figura 38). Este hecho, aunque no afectó considerablemente al entrenamiento, sí que se detectó en el proceso de transcripción, ya que el número de palabras obtenidas con el mismo número de hojas fue mayor.

255 Que copia su *mise en page*, manteniendo completamente la distribución de estrofas, para lo cual juega con los espacios interestróficos

Figura 37. Estrofas completas en la plana. (Biblioteca Nacional de España, INC/2159, h. aj^{r} – 82IM)



¶ Vita xpi fecho por coplas
por frey yñigo de mendoça a pe
ticiõ dela muy vtuosa señora do
ña iuana de cartagena.
¶ Ynuocaciõ del actor

¶ Clara sol diuinal
la cerrada niebla obscura
q̃ enel linaie humanal
por la culpa paternal
desde el comiẽço nos dura
despierta la voluntad
enderesça la memoria
por q̃ syn cõtrariedad
a tu alta magestad
se cante diuina gl̃ia

¶ Hõlla grand cõpassyõ
aquel amor entrañal
que por nr̃a saluaçion
hizo sofrir tal passion
a tu fijo natural
aq̃lla bondad diuina
q̃le forçó a ser ombre
ẽmiende lo q̃ se inclina
enesta carne mesquina
a offẽder el tu nombre

¶ Prosigue
¶ Los altos merecimiẽtos
de aq̃lla virgen ymadre
plos asperos tormentos
q̃ sufren por ti contẽtos
los q̃ te tienen porpadre
y lauitoria famosa
de tus martires passados
me alcançé que laprosa
de tu vida glorioſa
escriua ẽ metros rimados

¶ Despide las musas poeticas
e inuoca las cristianas

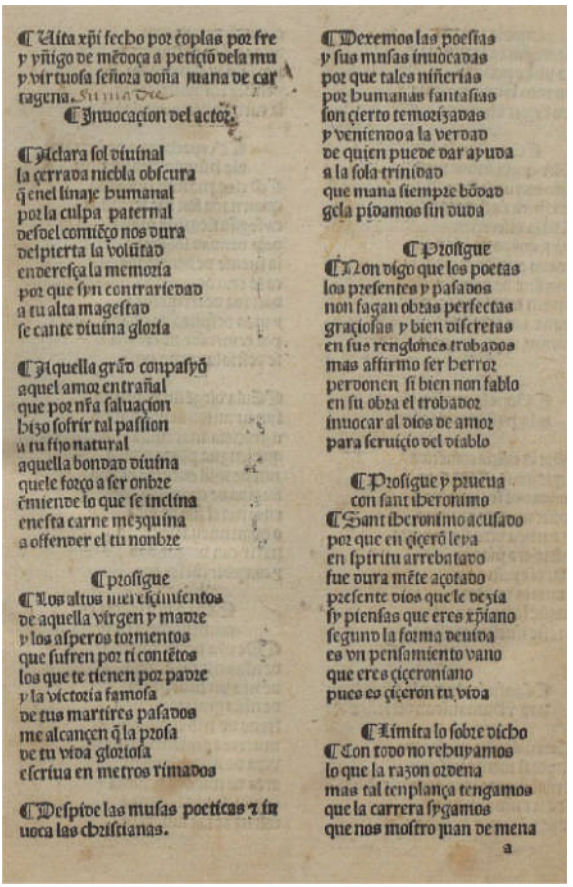
¶ Dexemos las poesias
y sus musas inuocadas
por q̃ tales ninirias
porhumanas fantasias
sõ cierto temorizados
y venjendo ala verdat
de q̃en puede dar aplaz
ala sola trinidat
q̃ mana sienpre bõdat
gela pidamos sin duda

¶ Prosigue
¶ Non digo q̃los poetas
los presentes y passados
non fagan obras pfectas
graçiosas y bien discretas
en sus rẽglones trobados
mas affirmo ser herror
perdonẽ sy bien nõ fablo
en su obra el trobador
inuocar al dios de amor
para seruiçio del diablo

¶ Prosigue y prueua
con sant iheronimo
¶ Sãt iheronimo acusado
por que en çiceron leya
en spiritu arrebatado
fue dura mente açotado
presente dios que le dezia
si pensas q̃ eres xpiano
segunt la forma deuida
es vn pensamiẽto vano
que eres ciceroniano
pues es çicerõ tu vida

a

Figura 38. Última estrofa de la plana incompleta (British Library, IB.52920, h. aj^{r} – 83*IM)



¶ Vita xpi fecho por coplas por fre
y pñigo de mẽdoça a petiçiõ dela mu
y virtuosa señora doña juana de car
tagena.

¶ Jnuocacion del actor.

¶ Aclara sol diuinal
la cerrada niebla obscura
q̃ enel linaje humanal
por la culpa paternal
desde el comiẽço nos dura
despierta la voluntad
endereça la memoria
por que sym contrariedad
a tu alta magestad
se cante diuina gloria

¶ Aquella grã compasyõ
aquel amor entrañal
que por nra saluaçion
hizo sofrir tal passion
a tu fijo natural
aquella bondad diuina
quele forçó a ser onbre
ẽmiende lo que se inclina
enesta carne mezquina
a offender el tu nonbre

¶ prosigue

¶ Los altos merescimientos
de aquella virgen y madre
y los asperos tormentos
que sufren por ti contẽtos
los que te tienen por padre
y la victoria famosa
de tus martires pasados
me alcançen q̃ la prosa
de tu vida gloriosa
escriua en metros rimados

¶ Despide las musas poeticas y in
uoca las christianas.

¶ Dexemos las poesias
y sus musas inuocadas
por que tales niñerias
por humanas fantasias
son cierto temorizadas
y veniendo a la verdad
de quien puede dar ayuda
a la sola trinidad
que mana siempre bõdad
gela pidamos sin duda

¶ Prosigue

¶ Non digo que los poetas
los presentes y pasados
non fagan obras perfectas
graciosas y bien discretas
en sus renglones trobados
mas affirmo ser herror
perdonen si bien non fablo
en su obra el trobador
inuocar al dios de amor
para seruiçio del diablo

¶ Prosigue y prueua
con sant iheronimo

¶ Sant iheronimo acusado
por que en cicerõ leya
en spiritu arrebatado
fue dura mẽte açotado
presente dios que le decia
sy pensas que eres xpiano
segund la forma dentro
es vn pensamiento vano
que eres ciceroniano
pues es ciceron tu vida

¶ Limita lo sobre dicho

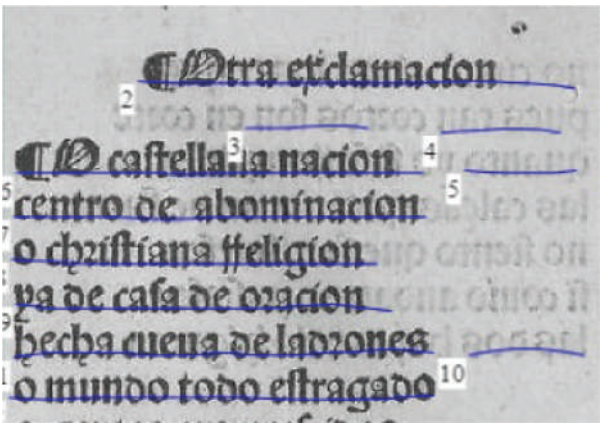
¶ Con todo no rebuyamos
lo que la razon ordena
mas tal templança tengamos
que la carrera sygamos
que nos mostro juan de mena

a

Al igual que se había hecho anteriormente, se corrigieron y adaptaron a las normas de edición las 25 planas transcritas, que se utilizaron para hacer evolucionar el modelo *Spanish Gothic Poetic Incunabula* a la siguiente versión 3, capaz de reconocer tres tipografías distintas y, en el caso de la que nos ocupa, con un CER del 1%. Esto nos viene a decir que, previsiblemente, cometerá un error por cada cien caracteres transcritos, lo que supone una tasa de éxito excelente. Este resultado, que mejora el que se había obtenido con el testimonio anterior, se debe principalmente al buen estado del ejemplar y a la nitidez de la tipografía utilizada.

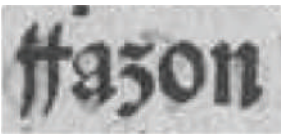
De la edición 90*IM únicamente se conserva un ejemplar mútilo en la Library of Congress de Washington, con diversos problemas añadidos a efectos de su transcripción automática. El primero y más evidente es la visualización del texto impreso de la plana opuesta, lo que conlleva que el algoritmo de segmentación interprete algunos renglones como líneas de texto y los marque como tal. Este comportamiento puede observase, en la imagen que produce como salida el algoritmo, en los puntos que identificó como líneas 3, 4, 5 y 10 del f. XVII^{r} (Figura 39), que hubo que eliminar manualmente.

Figura 39. Transparencia de la hoja y detección incorrecta de líneas (Library of Congress, Incun. X.M52 PQ6180, f. XVII^{r} – 90*IM)



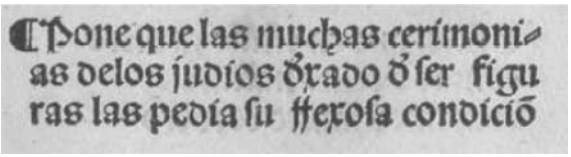
No es la única particularidad que presenta esta edición. Una característica tipográfica remarcable sería la morfología del dígrafo de la doble *r*, ya que, además de estar moldeada en un único tipo, se asemeja a una doble *f* (Figura 40). La interpretación de este tipo por parte del modelo de *Transkribus*, *Print M1*, fue inconsistente a lo largo del texto, transcribiéndolo como *ff*, *fr* y, en algunas ocasiones, de manera acertada, como *rr*.

Figura 40. Tipo de la grafía doble *r* aiiij^{r} (Library of Congress, Incun. X.M52 PQ6180, f. V^{r} – 90*IM)



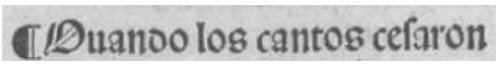
Asimismo, también destaca por la novedad, respecto a las anteriores ediciones, de incorporar el doble guion para separar palabras por cambio de renglón (Figura 41) —que el algoritmo transcribió como un guion simple[256]—, y, por primera vez también, la foliación en la parte superior derecha del recto de los folios, como un paratexto tipográfico que complica la aplicación del algoritmo y necesita marcarse manualmente para excluirlo en el proceso de segmentación.

Figura 41. Separación de sílabas por cambio de línea y abreviaturas (Library of Congress, Incun. X.M52 PQ6180, f. XV^{r} – 90*IM)



Los errores por confusión de grafía también existen en esta edición, un rasgo que, a pesar de no ser nuevo, sí que ofrece algunas peculiaridades aquí, como es el caso de la confusión de las mayúsculas *O* y *Q* que se produce al inicio de una estrofa del f. IX^{v} (Figura 42), cuya morfología es casi idéntica, a excepción del trazo del caído de la segunda de estas, muy pequeño y en horizontal a la línea de escritura (Figura 43).

Figura 42. Confusión de tipo (Library of Congress, Incun. X.M52 PQ6180, f. XVI^{r} – 90*IM)

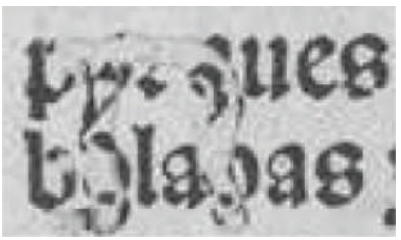


256 En 82*IM y en 99VC se utiliza, puntualmente, la barra inclinada para esta función.

Figura 43. Grafía de la letra *q* mayúscula (Library of Congress, Incun. X.M52 PQ6180, f. IX^{v} – 90*IM)

Más allá de las características de impresión, el ejemplar tiene agujeros provocados por xilófagos, algunos de ellos afectando al texto. Durante el proceso de digitalización, no se ha tenido la precaución de poner una hoja opaca en la parte trasera y, por tanto, se visualiza parcialmente contenido textual de la página siguiente (Figura 44). Esta circunstancia condicionará, en cierta medida, la interpretación que hará el modelo de transcripción automática, aunque no es un problema grave que haya afectado irremediablemente al ejemplar, sino algo puntual.

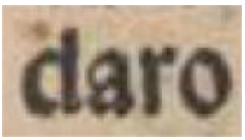
Figura 44. Agujero con texto de la página posterior visible (Library of Congress, Incun. X.M52 PQ6180, f. V^{r} – 90*IM)



Por otra parte, no presenta manchas de humedad que impidan la correcta visualización de la zona impresa, pero desafortunadamente, se han perdido los cuatro primeros folios y el décimo, por lo que el *Ground Truth* se generó a partir de las 25 primeras páginas conservadas. Con esta transcripción corregida, se volvió a entrenar al modelo *Spanish Gothic Poetic Incunabula* para crear su cuarta versión, esta vez, de nuevo con un CER del 1% para esta tipografía.

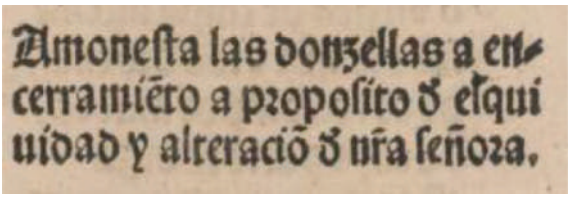
El ejemplar único de 91*VC, conservado en la BNE, está mútilo de tres hojas (h. aj-aij y a8) y, además, tiene otras tres desordenadas (h. a7, c8 y d5), colocadas, en esta secuencia, tras la h. diiij (Fernández Valladares, 2019, pp. 78-79). A esto se suman considerables roturas y manchas de humedad en diversas hojas. Del reconocimiento de su tipografía, cabe destacar que, cuando se combinan las letras *c* y *l* en lo que parece un mismo tipo (Figura 45), en el que quedan tan juntas, el modelo de *Transkribus*, *Print M1*, los interpretó como la letra *d*.

Figura 45. Combinación de las letras *c* y *l (*Biblioteca Nacional de España, INC/2900, h. aiij[r] – 91*VC)



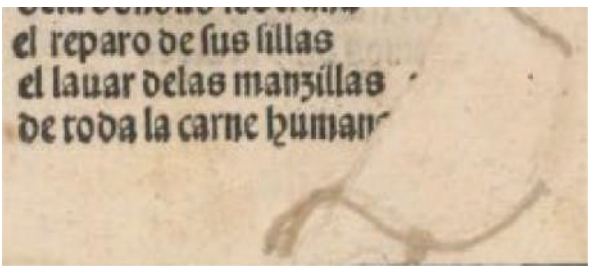
Aunque no incorpora foliación, como hacía 90*IM, sí que comparte el uso de la abreviatura para *de* y el doble guion para marcar la separación de sílabas por cambio de línea, que, lógicamente, se aplica solo a las rúbricas en prosa (Figura 46). Este carácter, con una forma muy similar a la del otro impreso, también se reconoció y transcribió automáticamente como un guion.

Figura 46. Separación de palabra para línea y abreviaturas (Biblioteca Nacional de España, INC/2900, h. aiij[r] – 91*VC)



Pese a las roturas y manchas de humedad que presenta en algunas hojas, el texto no se ve gravemente afectado, a excepción de una pequeña zona en la parte inferior de la h. biij que, además, ha sido pobremente restaurada (Figura 47).

Figura 47. Pérdida de texto (Biblioteca Nacional de España, INC/2900, h. biij[v] – 91*VC)



La creación del *Ground Truth* se hizo a partir de las 25 primeras planas del ejemplar y, con esta transcripción, se entrenó de nuevo el modelo *Spanish Gothic Poetic Incunabula*. Al igual que venía ocurriendo con los anteriores impresos, con esta muestra tipográfica se obtuvo, nuevamente, un CER del 1%, por lo que es previsible que el modelo resultante cometa un fallo por cada cien caracteres transcritos.

Dado que las otras dos ediciones de Pablo Hurus comparten tipografías (GW 2*134G/TW ma01246 y GW 3:100G/TW ma01247), se decidió utilizar únicamente una de ellas para efectuar el entrenamiento, en concreto, la de 1492. Como el ejemplar de París, además de estar mútilo, tiene graves manchas de humedad e, incluso, moho que afectan al texto[257], se escogió el ejemplar de la BNE para efectuar el entrenamiento.

Esta edición, al igual que 90*IM, está foliada y también utiliza el doble guion para separar las palabras truncadas por cambio de línea, de la misma manera que finaliza las rúbricas con un punto (Figura 48)[258], aunque por error también hay uno al final del primer verso de la primera estrofa (Figura 49).

Figura 48. Carácter de cambio de línea y uso de punto al final de la rúbrica (Biblioteca Nacional de España, INC/2900, f. II^{r} – 92VC)

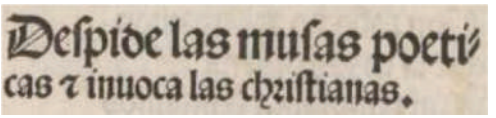


Figura 49. Punto en primer verso de la primera estrofa (Biblioteca Nacional de España, INC/2900, f. II^{r} – 92VC)

257 Tal y como destaca Massimo Marini en su descripción: "Presenta varias manchas de humedad y de moho, que no solo ha afectado el cartón debajo de la solapa en pergamino de la cubierta, sino también algunos folios del incunable, sobre todo los que han quedado del primer cuaderno. A causa de estas condiciones de conservación, en algunos puntos la tinta se encuentra muy descolorida, como es el caso del f. VI^{v}, lo que a veces dificulta la inteligibilidad de los versos o de las imágenes" (2023a).

258 Como ocurría con 82IM, aunque, en este caso, de manera aislada o como encontraremos en 99VC.

Sin embargo, por lo que destaca especialmente esta edición es por el uso de xilografías entremezcladas en el texto, cuya presencia provoca que el algoritmo de segmentación detecte algunas de sus zonas, por la forma del dibujo, como una línea de texto, como se puede apreciar en el número 9 de la Figura 50. Estos fallos se produjeron en varias ocasiones y hubo que corregirlos manualmente para evitar errores de transcripción.

Figura 50. Error en la detección de una línea de texto dentro de una xilografía (Biblioteca Nacional de España, INC/2900, f. III^{r} – 92VC)

No obstante, el mayor problema que tuvo el proceso de segmentación de las páginas que formaron el *Ground Truth* se produjo como consecuencia de la colocación del titulillo centrado en la parte superior del f. VIII, tanto en su recto como en su vuelto. Esta disposición, que extrañamente

no se repite en las páginas adyacentes pese a compartir el titulillo —donde se distribuye en dos secciones, cada una de las cuales encabeza una de las columnas (Figura 52)—, provocó que se considerase que ambas columnas formaban una única zona textual. En la Figura 51 se puede observar cómo la línea verde que marca la detección efectuada por el algoritmo, rodea ambas columnas, así como el titulillo y la foliación de la parte superior.

Figura 51. Titulillo centrado que causa confusión en detección de columnas (Biblioteca Nacional de España, INC/2900, f. VIII^{r} – 92VC)

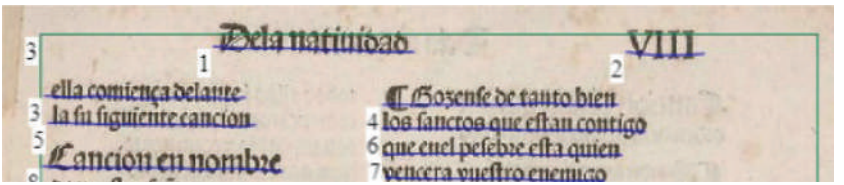
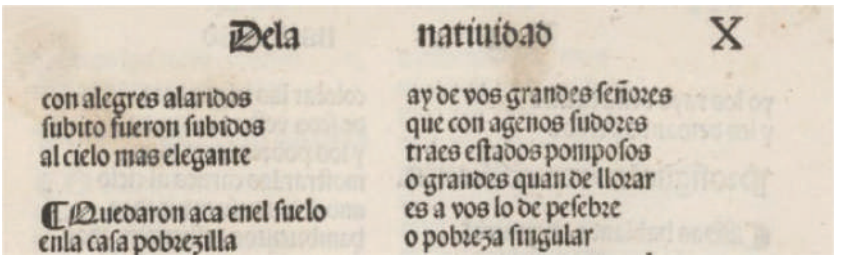
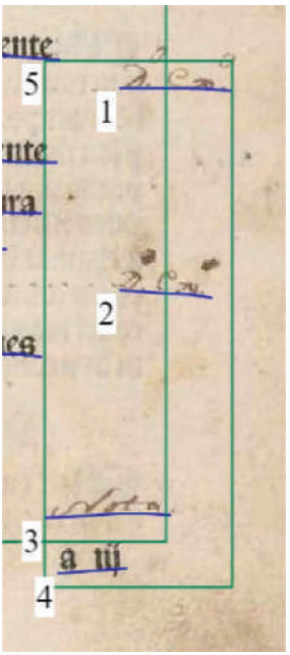


Figura 52. Titulillo distribuido en dos columnas (Biblioteca Nacional de España, INC/2900, f. X^{r} – 92VC)



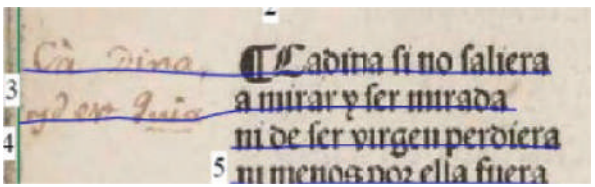
En cuanto al estado del ejemplar, este presenta múltiples anotaciones en los márgenes que dificultaron la detección correcta de las líneas. El algoritmo de segmentación se comportó de distinta manera según su separación y alineación respecto al texto principal. El caso más rápido de solucionar manualmente y totalmente predecible se producía cuando estas intervenciones manuscritas aparecían alejadas del texto impreso (Figura 53).

Figura 53. Anotaciones al margen (Biblioteca Nacional de España, INC/2900, f. III^{r} – 92VC)



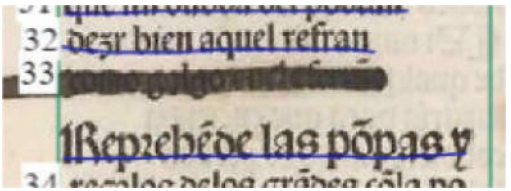
Pero, por otro lado, según se reducía esta distancia, la confusión del algoritmo aumentaba y, en diversas ocasiones, llegó a unir el paratexto con el texto principal al creer que pertenecía al mismo renglón (Figura 54). Corregir este comportamiento implicó eliminar de manera manual todos los puntos que formaban la línea establecida de manera automática por el proceso de segmentación.

Figura 54. Anotaciones detectadas como si fuesen la misma línea tirada (Biblioteca Nacional de España, INC/2900, f. III^{v} – 92VC)



El último detalle a destacar del ejemplar en cuanto a elementos que interfieren en el reconocimiento automático del texto es el tachado completo de una línea, "quizás por censura de algún lector o poseedor ante la comparación animal del refrán con el que fray Íñigo describe la condición del niño Jesús en el pesebre" (Marini, 2023a, p. 12). Aunque no fue detectada por el algoritmo de segmentación ni transcrita de forma automática (Figura 55), se incorporó manualmente el texto impreso original en el *Ground Truth*.

Figura 55. Tachado que impide detectar la línea de texto (Biblioteca Nacional de España, INC/2900, f. X^{r} – 92VC)



Con la segmentación y el correspondiente texto que formaban las páginas del *Ground Truth* corregidos, se generó la versión 6 del modelo *Spanish Gothic Poetic Incunabula*. Sorprendentemente, se obtuvo un CER del 0%, lo que indicaba que la transcripción de las páginas de prueba había sido perfecta. Pese a que no hay que olvidar que este número está basado en la probabilidad de que un evento ocurra y posiblemente no sea una tasa de error de cero absoluto, sí que nos asegura que las transcripciones de impresos con esta tipografía que obtendremos a partir de este modelo, aunque no serán necesariamente perfectas, sí que contendran pocos errores.

El último ejemplar que se utilizó para entrenar el modelo de transcripción, editado en 1499, es el que se encuentra en peor estado de conservación, hasta el punto de que, de las *Coplas de la vita Christi*, solo contamos con doce folios con fragmentos no consecutivos. El papel está en un estado muy frágil y las hojas han perdido la mayor parte del margen, aunque afortunadamente, la caja de texto no se ha visto afectada a excepción de la primera página, que tiene un agujero en la parte superior derecha. Como suele ser habitual en los procesos de digitalización, no se ha tenido la precaución de colocar una hoja en la parte posterior antes de tomar la foto y se visualiza nítidamente el texto de la página siguiente. Esta interferencia o contaminación textual, como ya se había adelantado, provocó que el algoritmo de segmentación creyese que los caracteres visibles a través del agujero pertenecían al mismo renglon que la página que estaba tratando.

Asimismo, al igual que ocurre con la edición 90*IM, se visualiza el texto impreso de la plana contraria, que, en algunos casos, se reconoció como zona textual susceptible de transcripción.

Al igual que las de Pablo Hurus de 1492 y 1495, esta edición incorpora titulillos y foliación impresos en su margen superior. La justificación centrada de los primeros, en este caso de forma sistemática, provocó el mismo resultado que en el f. VIII de 92VC, es decir, no se detectaron ambas columnas, por lo que hubo que segmentar manualmente cada una de las planas.

Pero lo más extraño, y que no se había producido anteriormente en ninguno de los ejemplares, fue la detección incorrecta de los renglones de texto. El algoritmo de segmentación tuvo un comportamiento caótico en algunas zonas, separando palabras de una misma línea o marcando únicamente la mitad del renglón. Aunque no se sabe con certeza qué produjo este fenómeno, la explicación más plausible es que interpretase algunas formas y combinaciones de letras como manchas.

El *Ground Truth* se elaboró a partir de las veinticuatro hojas que conserva el ejemplar y con él se entrenó la última versión del modelo *Spanish Gothic Poetic Incunabula*, la 7. Se obtuvo un CER del 1%, que se consideró muy bueno teniendo en cuenta el estado en el que se encuentra el ejemplar y el extraño comportamiento que había tenido el algoritmo de segmentación en la detección de las líneas de texto.

En resumen, se ha conseguido crear un modelo extendido con un CER o tasa de caracteres erróneos que se mueve entre el 0% y el 2%, dependiendo del ejemplar sobre el que se aplique, aunque en la mayoría de ocasiones está en torno al 1%, tal como se puede ver en la tabla adjunta. El mejor comportamiento lo presenta cuando se utiliza sobre 92VC, en la que da como resultado una transcripción prácticamente carente de fallos. En el peor de los casos, cuando se emplea sobre la *editio princeps*, obtiene un texto con el 98% de los caracteres correctos.

Tabla 5. Porcentajes de error

	82IM	**82*IM**	**83*IM**	**90*IM**	**91*VC**	**92VC**	**99VC**
CER	2%	0.7%	1%	1%	1%	0%	1%

En vista de estos resultados, es evidente que las transcripciones automáticas realizadas con el modelo *Spanish Gothic Poetic Incunabula* ofrecen textos fiables que permiten su utilización como punto de partida de una edición paleográfica y, con ello, explorar otros modelos complementarios de fijación textual.

6. Conclusiones

Las humanidades digitales nacen alrededor de la necesidad del tratamiento automático del lenguaje natural, estableciendo corpus de bases de datos sobre los cuales interactuar, lo que acaba derivando en múltiples aplicaciones, tanto lingüísticas, como literarias. Es, sobre todo, en el terreno de la lengua donde se ha implementado con gran rentabilidad el tratamiento digital de grandes corpus digitalizados, pero solo en los últimos años se han desarrollado aplicaciones similares centradas en la literatura, como es el caso de la estilometría o el análisis distante, que pueden ofrecer conclusiones de gran calado, que afectan a la propia atribución de autoría en relación a una obra anónima. Por otro lado, las diferentes bibliotecas nacionales e internacionales han hecho un gran esfuerzo de digitalización de sus fondos, que ha favorecido la deslocalización de la investigación, al menos en parte, para acceder hoy a través de un clic a materiales que se encuentran a cientos o miles de kilómetros del investigador en cuestión. Ahora bien, estos materiales no dejan de ser meras imágenes, en cuyo interior no se pueden hacer búsquedas si no se transforman en textos. Para todos estos desarrollos o aplicaciones digitales, partimos de una obviedad que no, por ello, es menos importante: si los corpus son la base de estos tratamientos, necesitamos disponer de estos a fin de entrenar cualquier software y/o tratarlos digitalmente, pero nos encontramos, a menudo, con cierta falta de disponibilidad o restricciones en su uso. Es aquí donde entra el reconocimiento automático de imágenes que agilice la obtención y fijación textual de estos materiales.

Si bien la función que ofrece *Gallica* para el reconocimiento textual de sus documentos digitalizados supone, sin lugar a dudas, un avance, al tratarse de un OCR convencional, no especializado en materiales antiguos, genera una tasa de éxito muy baja para textos con tipografía del siglo xv, con un índice de solo dos tercios de acierto, esto es, que, de cada cien grafías transcritas, aproximadamente treinta y tres presentarán errores. Es por esta razón que es necesario recurrir a un software de transcripción automática especializado en documentos antiguos, algo que solo ha sido posible

recientemente, gracias a la combinación de dos factores: la aparición de las redes neuronales que, si bien eran conocidas desde mediados del siglo xx, no se han podido desarrollar en cuanto a sus usos aplicados hasta alcanzar la velocidad de cálculo que ofrecen hoy en día los ordenadores actuales. Su aplicación al reconocimiento de imágenes con textos manuscritos e impresos de tipografías góticas medievales, mucho más inestables y complejas que la tipografía contemporánea, necesitaba de unos ordenadores de gran potencia para ejecutar en ellos softwares basados en este sistema de redes neuronales. Con ellas, podemos enseñar o, más bien, hacer aprender a los algoritmos, esto es, *entrenarlos*, para que los ingenieros informáticos no tengan que codificar la función que serán capaces de desarrollar por sí mismos. De esta manera, al aplicar esto al reconocimiento de imágenes y, en concreto, al reconocimiento de grafías, la consecuencia será que es el experto en la materia a tratar, en este caso los filólogos y/o los especialistas en bibliografía material, quien entrena el algoritmo y acaba haciendo uso de él, en una suerte de delegación de funciones del informático al humanista.

Ante tales necesidades, se han generado OCR que contemplan la transcripción automática de materiales antiguos mediante el entrenamiento de algoritmos basados en estas redes neuronales, entre los que destaca el paradigmático software de *Transkribus*, que, en un principio, se conocía como *Transcriptorium*, un proyecto financiado con fondos europeos para crear un sistema de uso abierto y cooperativo, que acabó convirtiéndose en un producto comercial. Ante esta nueva política de uso, que impedía su acceso de forma gratuita, surgieron otras alternativas de libre distribución, como es el caso de *eScriptorium*, basado en el motor de reconocimiento *Kraken*, tanto para la segmentación, como para la transcripción. *OCR4all*, una tercera plataforma similar a *Transkribus* y *eScriptorium*, solo utiliza *Kraken* para la primera de estas funciones, mientras que para la segunda emplea *Calamari*, también de libre distribución. La cohabitación de alternativas de software de transcripción ha favorecido la aparición de incipientes estudios comparativos de sus resultados, como el de Ayuso García (2022), derivado de la aplicación de *Transkribus* y *OCR4all* a un mismo conjunto de impresos de Arnao Guillén de Brocar, para obtener homogeneidad como corpus tipográfico, aunque desgraciadamente sus resultados no fueron concluyentes ni permitían decantarse por un software u otro.

En esta línea se enmarca el estudio comparativo de las tres plataformas de transcripción automática que se desarrolla en esta monografía en base a las 25 primeras planas del incunable zaragozano de Planck y Hurus de las *Coplas de la vita Christi* de fray Íñigo de Mendoza, desde de un análisis de sus respectivos motores de segmentación y de transcripción de la tipografía, algo necesario para elegir el software más adecuado para el establecimiento

de un modelo individual y es, a partir de él, que se generará un modelo extendido. Es el funcionamiento del algoritmo de segmentación el que acaba siendo decisivo, ahora sí, para la elección de un programa u otro, dado que *OCR4all*, a pesar de ofrecer una transcripción muy buena y ser un software en abierto, esto es, gratuito, presenta un proceso de segmentación de las páginas muy defectuoso y no permite entrenar la segmentación. Aunque el proyecto e*Scriptorium* sí que lo permite, utiliza el mismo software de segmentación que *OCR4all* y, por tanto, tiene sus mismas limitaciones, en términos de calidad, de lo que acaba derivándose en que los resultados del modelo de segmentación entrenado explícitamente en esta plataforma no son tan buenos como con el establecido por defecto por *Transkribus*. Si tenemos en cuenta que este último sí que permitía entrenar tanto un modelo de segmentación específico, como otro para la propia transcripción, y que los ofrecidos por defecto eran ya de una cierta calidad, tales posibilidades hacen decantarse por él, por *Transkribus*, a la hora de establecer una primera versión de nuestro modelo, individual y entrenado a partir del impreso 82*IM, al que se ha llamado *Spanish Gothic Poetic Incunabula*.

Sin embargo, los modelos individuales únicamente entrenados con un solo impreso presentan escasa rentabilidad, puesto que su excelente tasa de éxito se mantendrá solo si se aplica a incunables poéticos con esta misma tipografía, por lo que el siguiente paso sería establecer uno con mayor alcance, esto es, un modelo extendido. Los modelos sobreentrenados no parecen evidenciar, necesariamente, un aumento significativo de rendimiento, sino todo lo contrario, como ha puesto de manifiesto Fradejas Rueda al calificar los resultados de sus pruebas del reciente *Coloso español* a partir de algunos folios de manuscritos castellanos medievales como "muy pobres, cuando no directamente desastrosos" (2023, p. 16). El camino a seguir sería, por tanto, explorar modelos extendidos especializados en periodos temporales concretos, como aquel que desde el *Progetto Mambrino* se aplicó a los libros de caballerías del siglo xvi, después con un espectro de actuación más amplio aún; o como el aplicado por Gille Levenson (2023) mediante el uso de *eScriptorium*, generando un *dataset* a partir del incunable sevillano de Ungut y Polono de 1494 de la traducción castellana del *Regimiento de príncipes* de Egidio Romano y de otros diez manuscritos que se conservan de la misma obra, todo ello para generar un nuevo modelo extendido, cuya exactitud alcanzaba un 96.30% de éxito en su aplicación a un incunable y a manuscritos copiados entre el siglo xiii y el xv.

A la hora de establecer estos modelos extendidos más acotados y no sobredimensionados, con el propósito de afinar su especialización sin limitarla, se podría optar bien por un corpus formado por ediciones de distintas obras y de diversos impresores, o bien por una variedad de testimonios de

una misma obra. El objetivo de ambas perspectivas es lograr una variedad tipográfica que permita ampliar su campo de aplicación. De esta manera, el conjunto de letrerías que nos ofrecen las ocho ediciones incunables de las *Coplas de la vita Christi* de fray Íñigo de Mendoza, con hasta diez tipografías diferentes, siete de ellas para el texto en verso y las otras tres empleadas en sus rúbricas, singularizan este corpus poético impreso como el más adecuado del siglo xv para que, a partir de un entrenamiento diversificado, se aumente su grado de precisión y aplicabilidad. Para ello, se ha atendido no solo a la transcripción de la tipografía, sino también a las peculiaridades de la puesta en página de los propios incunables poéticos, que, normalmente, complican su segmentación, como son, entre otras y añadidas a las habituales de cualquier otro impreso, la coexistencia de texto centrado frente a la composición por columnas o, sobre todo, la falta de justificado derecho de la caja de escritura, por la inestable relación entre grafías y pies métricos, la abundancia de espacios interestróficos y otros blancos tipográficos que conllevan una visualización del texto de la plana contraria mayor que en ediciones en prosa. A pesar del entrenamiento de ambas redes neuronales, la de segmentación y la de transcripción, la intervención manual siempre será necesaria, desde la mera comprobación y confirmación, hasta actuaciones de mayor calado, como la segmentación prácticamente manual que han acabado requiriendo los incunables 82IM y 99VC, por los graves problemas materiales de los ejemplares o de su digitalización.

En definitiva, los resultados de esta monografía han permitido el desarrollo específico de un modelo extendido de una red neuronal de transcripción automática que sea aplicable con éxito a todos los incunables poéticos en tipografía gótica e, incluso, con un potencial que asegure una rentabilidad o eficacia de uso cercana al 100% en la transcripción automática de impresos similares del siglo xvi, con lo que ello implica como avance técnico y científico para el proceso de fijación textual de la poesía de cancionero. De momento, aquí queda para los estudios especializados en poesía impresa de cancionero y de romancero el modelo *Spanish Gothic Poetic Incunabula*, en su última versión extendida, que se ha puesto en acceso abierto para su uso gratuito por parte de la comunidad ciéntifica en la plataforma de *Transkribus*.

7. Índice de Figuras

8. Índice de Tablas

BIBLIOGRAFÍA

9. Bibliografía

Abadal, E. (2012). *Acceso abierto a la ciencia*, UOC.

Adamson, D. (1995). *Blaise Pascal: Mathematician, Physicist and Thinker about God*. Macmillan.

Agenjo, X. y Hernandez, F. (2020). OAI-PMH y Linked Open Data en el contexto de Hispana y Europeana: algunas reflexiones históricas. *Italian Journal of Library, Archives and Information Science*, *11*(1), 1-16. https://doi.org/10.4403/jlis.it-12573

Alarcón, E. (2002). El proyecto Corpus Thomisticum: descripción y perspectivas. *Anuario Filosófico*, *35*(3), 791-801.

Alberch, R. (2009). Archivos, la doble faz de la digitalización. En J. Vives (ed.), *Digitalización del patrimonio: archivos, bibliotecas y museos en la red* (pp. 25-86). UOC.

Allo, M. A. (1997). Teoría e historia de la conservación y restauración de documentos. *Revista General de Información y Documentación*, *7*(1), 253-295.

Alpaydin, E. (2014). *Introduction to Machine Learning*. The MIT Press.

Alvar Ezquerra, M. (1976). Obtención automática de índices de rimas y de sufijos. *Revista de dialectología y tradiciones populares*, *32*, 35-42.

Alvárez, J. y Vives, J. (2009). Las políticas internacionales de digitalización y su desarrollo en España. En J. Vives (ed.), *Digitalización del patrimonio: archivos, bibliotecas y museos en la red* (pp. 25-86). UOC.

Anglada, L. y Comellas, N. (2000). La Biblioteca Digital de Catalunya: oportunidades, opciones y estrategias en la adquisición compartida de información electrónica. En N. R. Brisaboa *et al.* (Eds,), *Primeras Jornadas de bibliotecas digitales* (pp. 237-248). Universidad de Valladolid.

Anglada, L. y Comellas, N. (2010). *Biblioteca Digital de Catalunya: 10 anys d'activitats* [Ponencia]. En 12 jornades catalanes d'informació i documentació, Barcelona, España.

Antonio, N. (1672). *Bibliotheca Hispana Nova sive Hispanorum qui usquam unquamve sive latina sive populari sive alia quavis lingua scripto aliquid consignaverun*. Nicolai Angeli Tinassii.

Antonio, N. (1788). *Bibliotheca Hispana Vetus, sive Hispani scriptores qui ab Octaviani Augusti aevo ad annum Christi md floruerunt*. Apud Viduam et Heredes D. Joachimi Ibarrae Regii Quondam Typographi.

Antonio, N. (1783-1788). *Bibliotheca Hispana Nova, sive hispanorum scriptorum qui ab anno MD. ad MDCLXXXIV floruere notitia*. Apud Joachimum de Ibarra - Apud Viudam et Heredes Joachimi de Ibarra.

Araújo, T. (3 de mayo de 2023). *Revisões literárias: a aplicação criativa de romances antigos (sécs. xv-xviii)* [*RELIT-Rom*]. Universidade Nova de Lisboa, https://relitrom.pt/

Archivo Digital del Romancero (29 de abril de 2023). Fundación Ramón Menéndez Pidal. [ADR] https://fundacionramonmenendezpidal.org/archivo-del-romancero/

Ariza, M. *et al.* (1973). Atlas lingüísticos plurilingües con ordenadores electrónicos. *Boletín del Centro de Cálculo de la Universidad de Madrid*, *23*, 12-15.

Arroyo Galán, L. (2005). *100 años de Informática y Telecomunicaciones. España siglo xx*. Fundación Rogelio Segovia para el Desarrollo de las Telecomunicaciones.

Askins, A. L-F. *et al.* (3 de febrero de 2023). *Bibliografia de Textos Antigos Galegos e Portugueses* [BITAGAP]. PhiloBiblon. Berkeley, The Bancroft Library - University of California Berkeley. https://bancroft.berkeley.edu/philobiblon/bitagap_es.html

Asunción, J. (2009). *El Papel: técnicas y métodos tradicionales de elaboración*. Parramón.

Avenoza, G. (2019). Codicología: estudio material del libro medieval. En G. Avenoza *et al.* (Eds.), *La producción del libro en la Edad Media* (pp. 57-130). Sílex.

Ayuso García, M. (2022). Las ediciones de Arnao Guillén de Brocar de BECLaR transcritas con ayuda de Transkribus y OCR4all: creación de un modelo para la red neuronal y posible explotación de los resultados. *Historias Fingidas*, *Numero Speciale 1: Digital Humanities e studi letterari ispanici*, 151-173.

Babbage, C. (1825). A note respecting the Application of Machinery to the Calculation of Astronomical Tables. En *Memoirs of the Royal Astronomical Society 1(II)* (p. 309). Royal Astronomical Society, Baldwin, Cradock and Joy.

Babbage, C. (1837). On the Mathematical Powers of the Calculating Engine. En B. Randell (ed.), *The Origins of Digital Computers* (pp. 19-54). Springer. https://doi.org/10.1007/978-3-642-61812-3_2

Babbage, C. (1864). *Passages from the Life of a Philosopher*. Longman, Roberts & Green.

Badia, L., Blecua, J. M., Claveria, G., Pujol, J., Soberanas, A. y Torruella, J. (1995). *Els cançoners catalans medievals. Concordances* [microfichas]. Fundació La Caixa - Seminari de Filologia i Informàtica, Universitat Autònoma de Barcelona.

Baird, H. S. (2014). A Brief History of Document and Writing Systems. En D. Doerman y K. Tombre (Eds.), *Handbook of Document Image Processing and Recognition* (pp. 3-10). Springer Reference.

Baird, H. S. y Tombre, K. (2014). The Evolution of Document Image Analysis. En D. Doerman y K. Tombre (Eds.), *Handbook of Document Image Processing and Recognition* (pp. 64-71). Springer Reference.

Balmaceda Abrate, J. C. (2008). Apuntes para el estudio del papel y las filigranas durante el siglo xv en la Corona de Aragón. *Aragón en la Edad Media*, *20*, 103-116.

Barbier, F. (2013). *Histoire des Bibliothèques. D'Alexandrie aux biblithèques virtuelles*. Armand Colin.

Barnes, S. B. (1997). Douglas Carl Engelbart: Developing the Underlying Concepts for Contemporary Computing. *IEEE Annals of the History of Computing*, *19*(3), 16-26. https://doi.org/10.1007/BF02404370

Bayerische Staatsbibliothek Inkunabelkatalog [*BSB-Ink*] (4 de mayo de 2023). Bayerische Staatsbibliothek. https://inkunabeln.digitale-sammlungen.de/sucheEin.htm

Bazzaco, S. (2018). El Progetto Mambrino y las tecnologías OCR: estado de la cuestión. *Historias Fingidas*, *6*, 257-272.

Bazzaco, S. (2020). El reconocimiento automático de textos en letra gótica del Siglo de Oro: creación de un modelo HTR basado en libros de caballerías del siglo xvi en la plataforma Transkribus. *Janus*, *9*, 534-561.

Bazzaco *et al.* (2022). Sistemas de reconocimiento de textos e impresos hispánicos de la Edad Moderna. La creación de unos modelos de HTR para la transcripción automatizada de documentos en gótica y redonda (s. xv-xvii). *Historias Fingidas*, *Numero Speciale 1: Digital Humanities e studi letterari ispanici*, 67-125.

Beltran, V. *et al.* (6 de julio de 2022). *Bibliografia de Textos Antics Catalans* [BITECA]. PhiloBiblon. Berkeley, The Bancroft Library - University of California Berkeley. http://sunsite.berkeley.edu/Philobiblon/phhmbi.html

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. En J. Kogan *et al.* (Eds.), *Grouping Multidimensional Data* (pp. 25-71). https://doi.org/10.1007/3-540-28349-8_2

Berners-Lee, T. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. HarperBusiness.

Biblioteca digital europea. (26 de julio de 2023). Comisión Europea y Fundación Europeana. [Europeana]. https://www.europeana.eu/

Biblioteca Patrimonial Digital de la Universitat de Barcelona [BiPaDi]. (26 de julio de 2023). Universitat de Barcelona. https://bipadi.ub.edu

Blasut, G. (2022). Los modelos de HTR *Silves1549_BNE* y *Spanish Gothic* como herramientas de la labor ecdótica. *Historias Fingidas*, *Numero Speciale 1: Digital Humanities e studi letterari ispanici*, 175-193.

Blei, D. M. *et al.* (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Bloy, C. H. (1967). *A History of Printing Ink, Balls, and Rollers 1440-1850*. Wynkyn De Worde Society.

Bodleian Incunables Catalog [*Bod-Inc online*]. (4 de mayo de 2023). Bodleian Libraries. http://incunables.bodleian.ox.ac.uk

Booth, A. D. y Locke, W. N. (1955). Historical Introduction. En A. D. Booth y W. N. Locke (Eds.), *Machine Translation of languages: fourteen essays* (pp. 1-14). Praeger Publishers.

Bordier, C. (2023). Gallica: Bibliothèque nationale de France (BnF). *American Jornalism*, *40*(1), 131-132.

Bouyer, C. (1994). *L'historie du papier*. Brepols.

Brea, M. y Lorenzo, P. (3 de mayo de 2023). *Base de datos da lírica galego-portuguesa* [*MedDB*]. Centro Ramón Piñeiro para a Investigación en Humanidades. http://bernal.cirp.gal/ords/f?p=MEDDB3:2

Brea, M. y Lorenzo, P. (3 de mayo de 2023). *Base de datos paleográfica da lírica galego-portuguesa* [*Palmed*]. Centro Ramón Piñeiro para a Investigación en Humanidades. http://bernal.cirp.gal/ords/palmed/r/palmed/inicio

Breuel, T. M. (2017). *High Performance Text Recognition using a Hybrid Convolutional-LSTM Implementation* [Ponencia]. En 14th International Conference on Document Analysis and Recognition, Kyoto, Japón. https://doi.org/10.1109/ICDAR.2017.12

Breuel, T. M. *et al.* (2013). *High-Performance OCR for Printed English and Fraktur using LSTM Networks* [Ponencia]. En 12th International Conference on Document Analysis and Recognition, Washington D.C., Estados Unidos. https://doi.org/10.1109/ICDAR.2013.140

Briquet, C. (1907). *Les filigranes. Dictionnaire historique des marques du papier*. W. Künkig & Fils.

Buckland, M. K. (1992). Emanuel Goldberg, Electronic Document Retrieval, and Vannevar Bush's Memex. *Jornal of the American Society for Information Science*, *43*(4), 284-294.

Buckland, M. K. (2006). *Emanuel Goldberg and his knowledge machine: information, invention, and political forces*. Libraries Unlimited.

Burrows, J. (2006). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing*, *22*, 27-47. https://doi.org/10.1007/BF02404370

Burton, D. M. (1981). Automated concordances and word indexes: The fifties. *Computer and Humanities*, *15*, 1-14. https://doi.org/10.1007/BF02404370

Busa, R. (1974-1980). *Index Thomisticus*. Frommann-Holzboog.

Busa, R. (1980). The Annals of Humanities Computing: the Index Thomisticus. *Computer and Humanities*, *14*(2), 83-90. https://doi.org/10.1007/BF02403798

Bush, V. (1931). The differential analyzer. A new machine for solving differential equations. *Journal of the Franklin Institute*, *212*(4), 447-488. https://doi.org/10.1016/S0016-0032(31)90616-9

Bush, V. (1945). As We May Think. *Atlantic Monthly*, *176*(1), 101-108.

Cabré, M. y Martí, S. (2 de abril de 2023). *Poesia i cançoners catalans medieval* [*Cançoners DB*]. Universitat de Girona. https://candb.narpan.net

Cabré Castellví, M. T. (2019). Bernard Quemada, una figura cabdal en la lingüística francesa (1926-2018). *Estudis Romanics*, *41*, 691-725.

Campbell, M. *et al.* (2002). Deep Blue. *Artificial Intelligence*, *134*(1), 57-82. https://doi.org/10.1016/S0004-3702(01)00129-1

Campo, I. *et al.* (1973). El análisis sintáctico automático como una ayuda para la elaboración del diccionario. *Boletín del Centro de Cálculo de la Universidad de Madrid*, *23*, 16-31.

Candela, G. *et al.* (2018). Migration of a Library Catalogue into RDA Linked Open Data. *Semantic Web Journal*, *9*(4), 481-491. https://doi.org/10.3233/SW-170274

Canet, J. L. (2000). La nueva Biblioteca Digital de la Universitat de València. En N. R. Brisaboa *et al.* (Eds.), *Primeras Jornadas de bibliotecas digitales* (pp. 71-78). Universidad de Valladolid.

Canet, J. L. (2004). Literatura i impremta durant el segle xvi a València. En *Escriptors valencians de l'Edat Moderna* (pp. 19-32). Acadèmia Valenciana de la Llengua.

Canet, J. L. (2014). Reflexiones sobre las humanidades digitales. *Janus*, *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, *Anexo 1*, 11-20.

Canet, J. L. (2019). Tipobibliografía valenciana de los siglos xv y xvi. *Historias Fingidas*, *7*, 455-458. https://doi.org/10.13136/2284-2667/139

Canet, J. L. y Haro, M. (1 de abril de 2023). *Servidor Web de Literatura Española*. Universidad de Valencia. [*Parnaseo*] https://parnaseo.uv.es

Castaños Alés, E. (2000). *Los orígenes del arte cibernético en España: el seminario de Generación Automática de Formas Plásticas del Centro de Cálculo de la Universidad de Madrid: (1968-1973)*. Biblioteca Virtual Miguel de Cervantes.

Catálogo Colectivo del Patrimonio Bibliográfico Español [CCBP] (15 de abril de 2023). Ministerio de Cultura y Deporte Gobierno de España. http://catalogos.mecd.es/CCPB/cgi-ccpb/abnetopac

Centre for the Study of the Cantigas de Santa Maria [CSM]. Universidad de Oxford. https://csm.mml.ox.ac.uk

Cheeseman, P. y Stutz, J. (1996). Bayesian Classification (AutoClass): theory and results. *Advances in knowledge discovery and data mining*, *180,* 153-180.

Chiron, G. *et al.* (2017). Impact of OCR errors on the use of digital libraries: towards a better access to information. En *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries* (pp. 249-152). IEEE. https://doi.org/10.5555/3200334.3200364

Clevier, D. (1992). *AI: the tumultuous history of the search for artificial intelligence*. BasicBooks.

Cohen, B. (2000). Howard Aiken and the Dawn of the Computer Age. En R. Rojas y U. Hashagen (Eds), *The First Computers: History and Architectures* (pp. 107-120). MIT Press.

Comisión de las Comunidades Europeas (2005). *i2010: Bibliotecas digitales*. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52005DC0465

Concheff, B. J. (1985). *Bibliography of Old Catalan Texts* [BOOCT]. Hispanic Seminar of Medieval Studies.

Copeland, J. (2012). *Turing: Pioneer of the Information Age*. Oxford University Press.

Corpus de Referencia del Español Actual [CREA]. (15 de abril de 2023). Real Academia Española. https://corpus.rae.es/creanet.html

Corpus del Español del Siglo xxi [CORPES]. (15 de abril de 2023). Real Academia Española. https://apps2.rae.es/CORPES/

Corpus Diacrónico del Español [CORDE]. (15 de abril de 2023). Real Academia Española. https://corpus.rae.es/cordenet.html

Corral, M. (2009). *La Biblioteca Digital Hispánica* [Ponencia]. En IX Workshop REBIUN sobre Proyectos Digitales, Salamanca, España.

Craine, A. G. (2022). *Ray Kruzwell*. Encyclopedia Britannica. https://www.britannica.com/biography/Raymond-Kurzweil

Crespo Nogueira, C. (1992). El papel soporte gráfico desde la Edad Media a la época actual. En *El papel y las tintas en la transmisión de información: primeras jornadas archivísticas* (pp. 193-202). Diputación de Huelva.

Crosas, F. (2024). Edición crítica del *Sermón trobado* de fray Íñigo de Mendoza. *Estudios Románicos*, *33*, 21-52.

Cultura i literatura de la baixa edat mitjana (25 de abril de 2023). Universitat Autònoma de Barcelona, Universitat de Barcelona i Universitat de Girona. [NARPAN] https://narpan.net

Daelemans, W. (2013). Explanation in Computational Stylometry. En A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing 2013* (pp. 451-462). Springer. https://doi.org/10.1007/978-3-642-37256-8_37

Dahl, S. (1982). *Historia del libro*. Alianza Universidad.

De Beni, M. (2023a). Caracterización lingüística de la *editio princeps* de las *Coplas de vita Christi* de Íñigo de Mendoza (POECIM/82IM). En J. Ll. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim82im

De Beni, M. (2023b). Caracterización lingüística de la *editio princeps* de las *Coplas de vita Christi* de Íñigo de Mendoza (POECIM/83*IM). En J. Ll. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim83im

De Beni, M. (2024). La caracterización lingüística de los incunables poéticos de Antón de Centenera. *Estudios Románicos*, *33*, 53-68.

De Beni, M. (en prensa). La lengua de los incunables poéticos zaragozanos de *Las coplas de vita Christi* de fray Íñigo de Mendoza. *Scripta. Revista Internacional de Literatura i Cultura Medieval i Moderna*, *24*.

De Ceglia, R. *et al.* (2023). Specialized astrocytes mediate glutamatergic gliotransmission in the CNS. *Nature*, *622*, 120-129. https://doi.org/10.1038/s41586-023-06502-w

De Sousa Neto, A. *et al.* (2020). HTR-Flor: A Deep Learning System for Offline Handwritten Text Recognition. En *33rd SIBGRAPI Conference on Graphics, Patterns and Images* (pp. 54-61). SIBGRAPI.

Deyermond, A. (1997). La *Celestina* como cancionero. En J. L. Canet Vallés y R. Beltrán Llavador (eds.), *Cinco siglos de "Celestina": aportaciones interpretativas* (pp. 91-106). Universitat de València.

Díaz de Miranda, M. D. (2014). Mètodes de reproducció d'imatge de la filigrana. *Unicum*, *13*, 186-191.

Díaz Hidalgo, R. J. *et al.* (2018). New insights into iron-gall inks through the use of historically accurate reconstruccions. *Heritage Science*, *6*(63), 1-15. https://doi.org/10.1186/s40494-018-0228-8

Díez Garretas, M. J. (2010). El cancionero MN46 (BNE, ms. 18183): del impreso al manuscrito. En José Manuel Fradejas *et al.* (coord.) *Actas del XIII Congreso Internacional de la Asociación Hispánica de Literatura Medieval* (pp. 697-717). Asociación Hispánica de Literatura Medieval.

Díez Garretas, M. J. (2011). El cancionero ML1, copia manuscrita de un impreso: las *Coplas de la vita Christi* de fray Íñigo de Mendoza. En Josep Lluís Martos (coord.), *Del impreso al manuscrito en los cancioneros* (pp. 73-112). Centro de Estudios Cervantinos.

Díez Garretas, M. J., Martos, J. L. y Moreno M. (2012). Base de datos del "Cancionero general del siglo xv" (MN13). *CIM: Cancioneros Impresos y Manuscritos*. https://cancioneros.org/node/11730.

Directorio nacional de recursos digitales (26 de julio de 2023). Ministerio de Cultura y Deporte [Hispana]. https://hispana.mcu.es/

Doermann, D. y Tombre K. (eds.) (2014). *Handbook of Document Image Processing and Recognition*. Springer Reference.

Donadío Maggi de Gandolfi, M. C. (1992). Thomae Aquinatis: opera omnia cum hypertextibus in cd-rom. *Sapientia*, *XLVII* (185), 233-234.

Drucker, H. *et al.* (1999). Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks*, *10*(5), 1048-1054.

Dudczak, A. *et al.* (2012). Creation of Textual Versions of Historical Documents from Polish Digital Libraries. En P. Zaphiris *et al.* (eds.), *Theory and Practice of Digital Libraries* (pp. 89-94). Springer. https://doi.org/10.1007/978-3-642-33290-6_10

Dureau, J. y Clements, D. W. G. (1986). *Principios para la preservación y conservación de los materiales bibliográficos*. Dirección General del Libro y Bibliotecas.

Dutton, B. (1990-1991). *El Cancionero del siglo xv (c. 1360-1520)*. Universidad de Salamanca.

Eder, M. *et al.* (2016). Stylometry with R: a Package for Computational Text Analysis. *R Journal*, *8*(1), 107-121.

Ellison, J. W. (1956). *Nelson's Complete Concordance of the Revised Standard Version of the Bible*. Thomas Nelson & Sons.

Essinger, J. (2015). *El Algoritmo de Ada: la vida de Ada Lovelace, hija de lord Byron y pionera de la era informática*. Alba.

Faulhaber, C. B. (1984). *Bibliography of Old Spanish Texts* (3ª ed.) [BOOST]. Hispanic Seminary of Medieval Studies.

Faulhaber, C. B. (2009). *PhiloBiblon*: pasado y futuro. *Incipit*, *XXIX*, 191-200.

Faulhaber, C. B. (2014). *PhiloBiblon*, Information Technology, and Medieval Spanish Literature: A Balance Sheet. En L. Soriano *et al.* (eds.), *Humanistats a la xarxa: món medieval* (pp. 15-43). Peter Lang.

Faulhaber, C. B. (2022). *PhiloBiblon* and the Wiki World. *Magnificat Cultura I Literatura Medievals*, *9*(0), 183-198.

Faulhaber, C. B. *et al.* (29 de diciembre de 2022). *Biblioteca Española de Textos Antiguos* [BETA]. PhiloBiblon. The Bancroft Library - University of California Berkeley. http://sunsite.berkeley.edu/Philobiblon/phhmbe.html

Fernández Valladares, M. (2019). Dos ejemplares recuperados del *Cancionero de Zaragoza* (92VC) con sorpresa inserta: unas desconocidas *Coplas del Quicumque vult* y dos nuevos fragmentos de *La Pasión trovada* y de la *Vita Christi*. *Revista de Cancioneros Impresos y Manuscritos*, *8*, 50-106.

Ferré, P. (25 de abril de 2023). *O Arquivo do Romanceiro Tradicional em Língua Portuguesa* [*Romanciero.pt*]. Universidade Nova de Lisboa. https://romanceiro.pt

Ferreiro, M. (2019). O portal *Universo Cantigas*: antecedentes, desenvolvemento e dificultades. En I. Tomassetti *et al.* (coord.), *Avatares y perspectivas del medievalismo ibérico vol. II* (pp. 1633-1644). Cilengua.

Ferreiro, M. (3 de mayo de 2023). *Universo Cantigas* [UC]. Universidade da Coruña. https://universocantigas.gal

Ferreras Fernández, T. (2018). Los repositorios instituciones: evolución y situación actual en España. En J. A. Merlo Vega (ed.), *Ecosistemas del Acceso Abierto* (pp. 39-84). Universidad de Salamanca.

Fisher, F. *et al.* (1983). *IBM and the U.S. Data Processing Industry: An Economic History*. Praeger.

Forniés Matías, Z. y García Quiroga, R. (2014). Factors de degradació intrínsecs als llibres: la naturalesa del material bibliogràfic. *BiD: textos universitaris de biblioteconomía i documentació*, *32*.

Fradejas Rueda, J. M. (2023). El Coloso español. *7 Partidas Digital*. https://7partidas.hypotheses.org/11531

Francis, W. N. (1965). A Standard Corpus of Edited Present-Day American English. *College English*, *26*(4), 267-273.

Francis, W. N. y Kucera, H. (1979). *Brown corpus manual*. Brown University.

Garcia, M. (1990). Introducción. En D. S. Severin (ed.), *El cancionero de Oñate-Castañeda* (pp. VII-XXXIV). Seminary of Hispanic Medieval Studies.

García Camarero, E. (1976). Noticias. *Boletín del Centro de Cálculo de la Universidad Complutense*, *28*, 102.

Garijo, F. J. y Verdejo, M. F. (1973). Preprocesador de Algol 60 por el método A.E.D. *Boletín del Centro de Cálculo de la Universidad de Madrid*, *22*, 48-62.

Garvin, M. (2022-2024). Repertorio abreviado *Fuentes impresas del Romancero (1501-1552)*. En J. L. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*, Alicante, Universitat d'Alacant. https://cancioneros.org/rar

Gesamtkatalog der Wiegendrucke [GW]. (4 de mayo de 2023). Staatsbibliothek zu Berlin. https://www.gesamtkatalogderwiegendrucke.de

Gille Levenson, M. (2023). Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR). *Journal of Data Mining and Digital Humanities*, *Special Issue: Historical documents and automatic text recognition*. https://doi.org/10.5281/zenodo.7387376

Girolamo, C. (25 de abril de 2023). *Repertorio informatizzato dell'antica letteratura trobadorica e occitana* [*TrobEU*]. Università di Napoli Federico II. http://www.rialto.unina.it

Girolamo, C. (25 de mayo de 2023). *Repertorio informatizzato dell'antica letteratura catalana* [RIALC]. Università di Napoli Federico II. http://www.rialc.unina.it

Girolamo, C. (25 de mayo de 2023). *Repertorio informatizzato dell'antica letteratura trobadorica e occitana* [RIALTO]. Università di Napoli Federico II. http://www.rialto.unina.it

Green, J. C. (1949). The Rapid Selector: An Automatic Library. *The Military Engineer*, *41*(283), 350-352.

Goff, F. R. (1973). *Incunabula in American Libraries. A Third Census of Fifteenth-Century Books Recorded in North American Collections*. Kraus Reprint Co.

Goldberg, E. (1932). Das Registrierproblem in der Photographie. En J. Eggert y A. von Biehler (eds.), *Bericht über den VIII Internationalen Kongreß für wissenschaftliche und angewandte Photographiehe* (pp. 317-320). Wiley. https://doi.org/10.1515/9783111718101.215

Goodfellow, I. *et al.* (2016). *Deep Learning*. The MIT Press.

Goodman, J. *et al.* (2007). Spam and the Ongoing Battle for the Inbox. *Communications of the ACM*, *50*(2), 25-33.

Gutiérrez Vázquez, S. (2012). Blaise Pascal: un matemático virtuoso. *Suma*, *Julio*, 105-114.

Ha, T. M., y Bunke H. (1997). Image Processing Methods for Document Image Analysis. En H. Bunke y P. S. P. Wang (eds.), *Handbook of Character Recognition and Document Image Analysis* (pp. 1-47). World Scientific.

Handel, P. W. (1933). *Statistical Machine* [Patente US1915993A]. General Electric Company.

Heide, L. (2009). *Punched-Card Systems and the Early Information Explosion 1880–1945*. The John Hopkins University Press.

Hernández Royo, P. (1994). *La imprenta valenciana de la familia Mey-Huete en el siglo xvi: producción y tipografía* [Tesis doctoral, Universitat de València]. Roderic.

Hidalgo Brinquis, M. C. (1992). Filigranas papeleras. En *El papel y las tintas en la transmisión de información* (pp. 193-202). Diputación de Huelva.

Hitzler, P. (2021). A Review of the Semantic Web Field. *Communications of the ACM*, *64*(2), 78-83. https://doi.org/10.1145/3397512

Hochreiter, S. y Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hollings, C. *et al.* (2019). *Ada Lovelace, la formación de una científica informática*. Publicacions de la Universitat de València.

Hutchins, J. (1999). The development and use of machine translation systems and computer-based translation tools [Ponencia]. En *International Symposium on Machine Translation and Computer Language Information Processing*, Beijing.

Iberian Books [IB]. (29 de abril de 2023). University College Dublin - School of History. https://iberian.ucd.ie

Incunabula Short Title Catalogue [ISTC]. (29 de abril de 2023). British Library. https://data.cerl.org/istc

Jacob, A. (2020). Punching Holes in the International Busa Machine Narrative. *IDEAH*, *1*(1).

Jockers, M. L. y Mimno D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769. https://doi.org/10.1016/j.poetic.2013.08.005

Jurafsky, D. y Martin, J. H. (2009). *Speech and Language Processing*. Pearson.

Keats, J. (2011). Del telar mecánico a la pianola y los primeros ordenadores. *Investigación y Ciencia*, *9*, 9.

Keefer, A. (2007). Los repositorios digitales universitarios y los autores. *Anales de Documentación*, *10*, 205-214.

Kiessling, B. *et al.* (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. En *2019 International Conference on Document Analysis and Recognition Workshops* (pp. 19-24). IEEE. https://doi.org/10.1109/ICDARW.2019.10032

Kim, Y. *et al.* (2020). Applying Computer Vision Systems to Historical Book Illustrations: Challenges and First Results. En S. Reinsone *et al.* (eds.), *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries* (pp. 255-260). RWTH Aachen University/CEUR Workshop Proceedings.

Kirchner, F. *et al.* (2016). OCR bei Inkunabeln – Offizinspezifischer Ansatz der Universitätsbibliothek Würzburg. *ABI Technik*, *36*(3), 178-188.

Koppel, M. *et al.* (2009). Computational Methods in Authorship Attribution. *JASIST*, *60*(1), 9-26.

Kraemer Koeller, G. (1973). *Tratado de la previsión del papel y de la conservación de bibliotecas y archivos*. Dirección General de Archivos y Bibliotecas.

Lacarra, M. J. (2019). El libro antiguo impreso. En Gemma Avenoza *et al.* (eds.), *La producción del libro en la Edad Media* (pp. 293-334). Sílex.

Lacarra, M. J. (4 de mayo de 2023). *Catálogo de obras medievales impresas en castellano* [COMEDIC]. Universidad de Zaragoza. https://comedic.unizar.es

Lacasta, J. *et al.* (2022). Tracing the origins of incunabula through the automatic identification of fonts in digitised documents. *Multimedia Tools and Applications*, *81*, 40977-40991. https://doi.org/10.1007/s11042-022-13108-3

Lafarga Coscojuela, M. A. (1994). *Biología celular de la neurona y de la sinapsis*. Universidad de Cantabria.

Leavitt, D. (2007). *Alan Turing: el hombre que sabía demasiado*. Antoni Bosch.

Lebert, M. (2005). Le Project Gutenberg (1971-2005) [Ponencia]. En *Troisième symposium international sur les études françaises valorisées par les technologies: langages et dialogues interculturels*, Toronto, Canada.

LeCun, Y. *et al.* (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*, 541-551.

LeCun, Y. *et al.* (1999). Object Recognition with Gradient-Based Learning. En *Shape, Contour and Grouping in Computer Vision. Lecture Notes in Computer Science, v. 1681* (pp. 319-345). Springer. https://doi.org/10.1007/3-540-46805-6_19

LeCun, Y. *et al.* (2010). Convolutional Networks and Applications in Vision. En *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (pp. 253-256). IEEE. https://doi.org/10.1109/ISCAS.2010.5537907

LeCun, Y. *et al.* (2015). Deep Learning. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Lee, S. D. (2001). *Digital Imaging, a practical handbook*. Neal-Schuman Publishers.

Lee, S. W. y Ryu, D. S. (2001). Parameter-Free Geometric Document Layout Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(11), 1240-1256.

Leibniz, G. (1703). Explication de l'arithmétique binaire. *Memoires de l'Académie Royale des Sciences*, *3*, 85-93.

Lesk, M. (1997). *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann Publishers.

Literatura Catalana de l'edat moderna (25 de abril de 2023). Universitat de Girona. [NISE] https://nise.cat

López, A. y Munarriz J. (2021). *El Centro de Cálculo de la Universidad de Madrid (1968-1973): ciencia, arte y creación computacional*. Ediciones Complutense.

López Casas, M. M. (2020). Materialidad y estructura de un temprano cancionero colectivo incunable (86*RL). *Revista de Poética Medieval*, *34*, 131-158.

López Casas, M. M. (2021). Los poemas de 86*RL, criterios de selección y relación con otros incunables poéticos: variación y variantes. *Criticón*, *141*, 133-156.

López Casas, M. M. y Mangas, N. A. (2022). El *Cancionero de Llabiá* (POECIM/86*RL). En J. L. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim86rl

López de Mántaras Badia, R. y Meseguer González, P (2017). *Inteligencia Artificial*. CSIC.

López Poza, S. (2020). Humanistas y Humanidades digitales. Trayectoria y proyección en la Filología española. En A. Egido *et al.* (eds.), *Humanidades y Humanismo. Homenaje a María Pilar Cuartero* (pp. 15-43). Institución Fernando el Católico.

López Poza, S. y Pena Sueiro, N. (1 de abril de 2023). *Biblioteca Digital Siglo de Oro* [BIDISO]. Universidade da Coruña. https://bidiso.es

Lu, Z. *et al.* (1999). Advances in the BBN BYBLOS OCR System. En *Proceedings of the Fifth International Conference on Document Analysis and Recognition* (pp. 337-340). IEEE. https://doi.org/10.1109/ICDAR.1999.791793

Lucía Megías, J. M. (2003). La "Informática Humanística": notas volanderas desde el ámbito hispánico. *Incipit*, *XXIII*, 91-114.

Lucía Megías, J. M. (2010). Los nuevos filólogos del siglo xxi: la literatura medieval hispánica en la red. En José Manuel Fradejas *et al.* (eds.), *Actas XIII Congreso AHML* (pp. 1233-1254). Ayuntamiento de Valladolid y Universidad de Valladolid.

Lucía Megías, J. M. (2012). *Elogio del texto digital*. Fórcola Ediciones.

Lupovici, C. *et al.* (2003). Les usages de Gallica. *Bulletin des Bibliothèques de France*, *48*(4), 40-44.

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL*, *226*, 1-7.

Madroñal, A. y Vega García-Luengos, G. (2021). *El renacer del Fénix: Yo he hecho lo que he podido, Fortuna lo que ha querido: Una nueva comedia de Lope de Vega*. Ediciones Universidad de Valladolid.

Marcos Marín, F., Faulhaber, C. y Gómez Moreno, A. (29 de diciembre de 2022). *Biblioteca electrónica Admyte. Archivo Digital de Manuscritos y Textos Españoles* [*Admyte*]. Hispanic Seminary of Medieval Studies, University of Wisconsin. https://www.admyte.com

Marcos Marín, F. A. (1971). Posibilidad y dificultades de la traducción automática. *Filología Moderna*, *42*, 313-327.

Marcos Marín, F. A. (1994). *Informática y humanidades*. Gredos.

Marcos Marín, F. A. (2009). Historia humana de la lengua española y su computación. *Studies in Hispanic and Lusophone Linguistics*, *2*(2), 387-416. https://doi.org/10.1515/SHLL-2009-1057

Marini, M. (2023a). *Vita Christi hecho por coplas* o *Cancionero de Zaragoza* (POECIM/92VC). En J. Ll. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim92vc

Marini, M. (2023b). *Vita Christi hecho por coplas* o *Cancionero de Zaragoza* (POECIM/95VC). En J. L. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim95vc

Marini, M. (2024). El *Cancionero de Zaragoza* (92VC y 95VC): estudio material e interno. *Estudios Románicos*, *33*, 187-204. https://doi.org/10.6018/ER.590161

Marini, M. (en prensa). Los ejemplares del *Cancionero de Zaragoza* (92VC y 95VC): materialidad e historia. *Magnificat*, *11*.

Martín Abad, J. (2010). *Catálogo bibliográfico de la colección de incunables de la Biblioteca Nacional de España*. Biblioteca Nacional de España.

Martín Pascual, L. (2020). Identidad y delimitación textual del poema 29 de Ausiàs March. *Estudios Románicos*, *29*, 319-330. https://doi.org/10.6018/ER.425131

Martín Pascual, L. (2024). La elaboración de una edición sinòptica de las poesías de Ausiàs March. *Specula. Revista de Humanidades y Espiritualidad*, *9*, 101-129. https://doi.org/10.46583/specula_2024.9.1119

Martínez-Conde, M. L. (2012). El proyecto EuropeanaLocal: los contenidos regionales y locales en Europeana. En *VI Congreso Nacional de Bibliotecas públicas* (pp. 139-147). Ministerio de Educación, Cultura y Deporte.

Martos, J. L. (ed.) (2011a). *Del impreso al manuscrito en los cancioneros*. Centro de Estudios Cervantinos.

Martos, J. L. (2011b). Hacia un canon de transmisión textual del cancionero medieval: del impreso al manuscrito. En Josep Lluís Martos (ed.), *Del impreso al manuscrito en los cancioneros* (pp. 207-212). Centro de Estudios Cervantinos.

Martos, J. L. (2012a). La Real Academia Española y el *Cancionero general del siglo* xv: un proyecto editorial ilustrado. *Boletín de la Real Academia Española*, *92*(306, julio-diciembre), 221-253.

Martos, J. L. (2012b). Josep Maria Torres Belda i la copia vuitcentista del *Cançoner de Saragossa*. En *Estudis de Llengua i Literatura Catalanes*, *v. 64*, *Miscel·lània Albert Hauf*, *3* (pp. 125-152). Publicacions de l'Abadia de Montserrat.

Martos, J. L. (2014). Ausiàs March en Italia: variantes y contextos de un *codex descriptus*. *Revista de Poética Medieval*, *28*, 265-294.

Martos, J. L. (2016). La Suplicació de natura humana de Joan Roís de Corella: fragmentos recuperados de una obra perdida. *Cultura neolatina*, *1-2*, 165-201.

Martos, J. L. (2018a). CIM: un espacio digital para la poesía de cancionero. Bases de datos. En G. Lalomia y D. Santonocito (eds.), *Literatura medieval (hispánica): nuevos enfoques metodológicos y críticos* (pp. 323-337). Cilengua.

Martos, J. L. (2018b). Fuentes poéticas incunables: el cancionero 87FD y Juan Tallante. En A. Zinato y P. Bellomi (eds.), *Poesía, poéticas y cultura literaria* (pp. 523-533). Ibis.

Martos, J. L. (2019). Ediciones valencianas en la imprenta incunable de Venecia. *Anuario de estudios medievales*, *49*(2, julio-diciembre), 683-704.

Martos, J. L. (2021). Manuscritos e incunables en el entorno de los Reyes Católicos: el cancionero EM6. *RILCE*, *37*(1), 319-346. https://doi.org/10.15581/008.37.1.319-46

Martos, J. L. (coord.) (2022a). *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim

Martos, J. L. (2022b). *Les trobes en llaors de la Verge Maria* (POECIM/74*LV). En J. L. Martos (coord.), *POECIM: Poesía, Ecdótica e Imprenta*. Universitat d'Alacant. https://cancioneros.org/poecim/poecim74lv

Martos, J. L. (25 de abril de 2023). *Cancioneros Impresos y Manuscritos* [CIM]. Universidad de Alicante. https://cancioneros.org

Martos, J. L. (2023a). *El primer cancionero impreso y un pliego poético incunable*. Iberoamericana - Vervuert.

Martos, J. L. (2023b). Las poesías en castellano de *Les trobes en lahors de la Verge Maria* (74*LV). *Revista de Filología Románica*, *40*, 135-151.

Martos, J. L. (2024a). Poesía de cancionero y fuentes impresas: el repertorio abreviado de incunables poéticos. En G. Vallín y M. I. Toro (eds.), Tradiciones poéticas de la Romania (entre la Edad Media y la Edad Moderna) (pp. 329-340). Universidad de Salamanca.

Martos, J. L. (2024b). Poemas y testimonios de Juan Tallante: delimitación y ampliación del corpus. *Estudios Románicos*, *33*, 205-252. https://doi.org/10.6018/ER.587471

Martos, J. L. (2024c). Tres ejemplares del *Cancionero de Llabià*. En A. I. Peirats (ed.), *Cultura i transmissió textual en l'occident europeu (segles xiv-xv)* (pp. 163-188). Tirant Humanidades.

Massoli, M. (1977). *Coplas de Vita Christi*. D'Anna.

Material Evidence in Incunabula [MEI]. (1 de mayo de 2023). https://data.cerl.org/mei

Mauchly, J. W. (1982). The Use of High-Speed Vacuum Tube Devices for Calculating [Reproducción de mecanuscrito no publicado de 1942]. En B. Randell (ed.), *The Origins of Digital Computers: Selected Papers* (pp. 355-258), Springer. https://doi.org/10.1007/978-3-642-61812-3_28

McCulloch, W. S. y Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*, 115-133.

McDowell, C. S. (2007). Evaluating Institutional Repository Deployment in American Academe Since Early 2005: Repositories by the Numbers, Part 2. *D-Lib Magazine*, *13*(9). https://doi.org/10.1045/september2007-mcdowell

Melero, R. (2008). El paisaje de los repositorios institucionales *open access* en España. *BiD: textos universitaris de biblioteconomía i documentació*, *20*.

Menabrea, L. F. (1843). Sketch of the Analytical Engine invented by Charles Babbage. En *Scientific Memoirs v. 3,* (pp. 666-731). Richard Taylor. [1ª ed. en francés: Notions sur la machine analytique de M. Charles Babbage. En *Bibliothèque Universelle de Genève v. 41* (pp. 352-376)].

Mendoza, I. (1953). *Vita Christi fecho por coplas de fray Íñigo de Mendoça* [facsímil]. Real Academia Española.

Mendoza, I. (1975). *Vita Christi fecho por coplas de fray Íñigo de Mendoça* [facsímil]. Antonio Pérez Gómez.

Meya, M. (1983). Editorial, *Procesamiento del Lenguaje Natural*, *1*, 1-4.

Millás Mascarós, E. y Escriche Soriano, M. (2017). La Biblioteca Digital de la Universitat de València. difusión y preservación de fondos históricos. *RUIDERAe: Revista de Unidades de Información*, *12*.

Mommsen, H. *et al.* (1996). X-Ray Fluorescence Analysis with Synchrotron Radiation on the Inks and Papers of Incunabula. *Archaeometry*, *38*(2), 347-357.

Moreno, M. (2011). Pliegos sueltos poéticos en cancioneros manuscritos: el *Cancionero Capitular de la Colombina* (SV2). En J. Ll. Martos (coord. y ed.), *Del impreso al manuscrito en los cancioneros* (pp. 47-71). Centro de Estudios Cervantinos.

Moreno, M. (2012a). Descripción codicológica MN13 (Mss. 3755-3765, Biblioteca Nacional, Madrid). En J. L. Martos (coord.), *Cancioneros Impresos y Manuscritos*. https://cancioneros.org/sites/default/files/2022-07/MN13.pdf

Moreno, M. (2012b). Inventario de fuentes manuscritas e impresas utilizadas en la formación de Proyecto de cancionero (MN13). En J. Ll. Martos (coord.), *Cancioneros Impresos y Manuscritos*. https://cancioneros.org/sites/default/files/2022-07/Fuentes%20de%20MN13.pdf

Murphy, K. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Navarro Colorado, B. (2019). Por un análisis distante y profundo: un corpus piloto de la poesía lírica castellana del Siglo de Oro. *Revista de poética medieval*, *33*, 51-76.

Nelson, T. H. (1965). A File Structure for The Complex, The Changing and the Indeterminate. En *Proceedings of the 1965 20th National Conference* (pp. 84-100). ACM. https://doi.org/10.1145/800197.806036

Neri, S. (2019). Proyecto Mambrino. *Historias Fingidas*, *7*, 443-448.

Neumann, J. (1945). *First draft of a report on the EDVAC* [documento inédito]. Moore School of Electrical Engineering, University of Pennsylvania.

Nikolaidou, K. *et al.* (2022). A survey of historical document image datasets". *International Journal on Document Analysis and Recognition*, *25*(4), 305-338.

Nilsson, N. J. (2010). *The Quest for Artificial Intelligence: a History of Ideas and Achievements*. Cambridge University Press.

Ning, L. (1991). *Report of An Implementation of OCR Based on Sketeleton Matching* [documento inédito]. University of Kent at Canterbury.

Nitti, J. J. (1975). Computers and the Old Spanish Dictionary. *Computers and the Humanities*, *12*(1-2), 43-52.

Nockels, J. *et al.* (2022). Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research. *Archival Science*, *22*, 367-392. https://doi.org/10.1007/s10502-022-09397-0

Norton, F. J. (1978). *A descriptive catalogue of printing in Spain and Portugal 1501-1520*. Cambridge University Press.

Nyhan, J. y Terras, M. (2017). Uncovering 'hidden' contributions to the history of Digital Humanities: the Index Thomisticus' female keypunch operators [Ponencia]. En *Digital Humanities Conference 2017*, Montreal.

O'Regan, G. (2016). *Introduction to the History of Computing*. Springer.

O'Regan, G. (2021). *A Brief History of Computing* (3ª ed). Springer.

Osareh, A. y Shadgar, B. (2010). Machine Learning Techniques to Diagnose Breast Cancer. En *2010 5th International Symposium on Health Informatics and Bioinformatics* (pp. 114-120). IEEE. https://doi.org/10.1109/HIBIT.2010.5478895

Palau y Dulcet, A. (1948-1977). *Manual del librero hispanoamericano*. Librería Palau.

Parkinson, S. (2019). Supporting Research with Poetic Metadata: The Oxford *Cantigas de Santa Maria* Database. *Revista de Poética medieval*, *33*, 77-86.

Passarotti, M. (2019). The Project of the Index Thomisticus Treebank. En M. Berti (ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital* (pp. 299-319). De Gruyter.

Patrimonio Digital Complutense [PDC]. (10 de agosto de 2023). Biblioteca Universidad Complutense. https://patrimoniodigital.ucm.es

Pêcheux, M. (1978). *Hacia el análisis automático del discurso*. Gredos.

Pedraza Gracia, M. J. (1997). *La producción y distribución del libro en Zaragoza 1501-1521*. Institución Fernando el Católico.

Pellón, I. *et al.* (2004). De la tinta china al tóner. Evolución de una técnica ancestral: la fabricación del *negro humo*. *Anales de la Real Sociedad Española de Química*, *Oct-Dic*, 45-54.

Pena Sueiro, N. (2017). El portal BIDISO: pasado, presente y futuro inmediato. Un ejemplo de evolución en aplicaciones de las HD. *Studia Aurea*, *11*, 73-93. https://doi.org/10.5565/rev/studiaaurea.264

Pena Sueiro, N. y Álvarez García, S. (2014). El *Catálogo y Biblioteca digital de relaciones de sucesos*: bases de datos bibliográficas, textos e imágenes. *Janus*, *Anexo 1*, *Humanidades Digitales: desafíos, logros y perspectivas de futuro*, 285-303.

Pérez Gómez, A. (1959). Notas para la bibliografía de Fray Íñigo de Mendoza y de Jorge Manrique, *Hispanic Review*, *27*(1), 30-41.

Pérez Priego, M. A. (2011). *La edición de textos* (2º ed.). Editorial Síntesis.

Petersen, S. H. (29 de abril de 2023). *Pan-Hispanic Ballad Project* [PHBP]. University of Washington. https://depts.washington.edu/hisprom/

Petrits, A. (2001). *EC Systran: the Commission's Machine Translation System*. European Commission Translation Service.

Pham, M. Q. *et al.* (2021). SYSTRAN @ WMT 2021: Terminology Task. En *Proceedings of the Sixth Conference on Machine Translation* (pp. 842-850). Association for Computational Linguistics.

Piccard, G. (1961-1997). *Die Wasserzeichenkartei Piccard im Hauptstaatsarchiv Stuttgart: Findbuch*. Kohlhammer.

Poon, J. C. H. *et al.* (1995). A robust vision system for vehicle licence plate recognition using grey-scale morphology. En *Proceeding of the IEEEE International Symposium on Industrial Electronics vol. 1* (pp. 394-399). IEEE.

Quintanar, J. L. (2010). *Neurofisiología básica*. Universidad Autónoma de Aguascalientes.

Ralph DiFranco, *et al.* (3 de febrero de 2023). *Bibliografía de Poesía Áurea* [BIPA]. PhiloBiblon. The Bancroft Library - University of California Berkeley. https://bancroft.berkeley.edu/philobiblon/bipa_es.html

Ralston, A. y Reilly, E. D. (1983). *Encyclopedia of Computer Science and Engineering* (2ª ed.). Van Nostrand Reinhold.

Reed, S. (2019). *Script, history and ideology: German fonts and handwriting*. British Library. https://blogs.bl.uk/european/2019/05/script-history-and-ideology.html

Rejewski, M. (1981). How Polish Mathematicians Deciphered the Enigma. *IEEE Annals of the History of Computing*. *3*(3), 213-234.

Repositori d'Objectes Digitals per a l'Ensenyament, la Recerca i la Cultura [RODERIC]. (15 de abril de 2023). Universitat de València. https://roderic.uv.es/

Reul, C. (2020). *An Intelligent Semi-Automatic Workflow for Optical Character Recognition of Historical Printing* [tesis doctoral, Bayerischen Julius-Maximilians-Universität Würzburg]. Universitätsbibliothek Würzburg.

Reul, C. *et al.* (2019). OCR4all: An Open-Source Tool Providing a (Semi-) Automatic OCR Workflow for Historical Printings. *Applied Sciences*, *9*(22). https://doi.org/10.3390/app9224853

Revina, I. M. y Emmanuel, W. R. S. (2021). A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University – Computer and Information Sciences*, *33*, 619-628.

Rivera, G. M. y Trienens, R. J. (1979). The *Cancionero de Íñigo de Mendoza*: An Unkown Fifteenth-Century Edition in the Library of Congress. *La Coronica*, *8*, 22-28.

Rodríguez Brisaboa, N. *et al.* (2019). Sagrario López Poza, humanista digital. En C. Fernández Travieso y N. Pena Sueiro (eds.), *Festina Lente. Augusta empresa correr a espacio. Studia in honorem Sagrario López Poza* (pp. 33-42). Servizo de Publiacións Universidade da Coruña.

Rodríguez Ferrer, R. (2007). *Materialidad y materiales del cancionero SA4 (ms. 2139 de la Biblioteca Universitaria de Salamanca)* [trabajo de grado, Universidad de Salamanca]. Universidad de Salamanca.

Rodríguez Puértolas, J. (ed.) (1968a). *Fray Íñigo de Mendoza y sus 'Coplas de Vita Christi'*. Gredos.

Rodríguez Puértolas, J. (ed.) (1968b). Fray Íñigo de Mendoza, *Cancionero*. Espasa-Calpe.

Rojas, R. (1997a). Los ordenadores de Konrad Zuse. *IEEE Annals of the History of Computing*, *19*(2), 5-16.

Rojas, R. (1997b). Konrad Zuse's Legacy: The Architecture of the Z1 and Z3. *Investigación y Ciencia*, *22*, 22-30.

Rojas, R. *et al.* (2005). The Reconstruction of Konrad Zuse's Z3. *IEEE Annals of the History of Computing*, *27*(3), 23-32.

Rojas-Sola, J. I. *et al.* (2021). Blaise Pascal's Mechanical Calculator: Geometric Modelling and Virtual Reconstruction. *Machines*, *9*(7), 136. https://doi.org/10.3390/machines9070136

Rojo, G. (2015). Corpus textuales del español. En J. Gutiérrez (ed.), *Enciclopedia de Lingüística Hispánica* (pp. 285-296). Routledge.

Rojo, G. (2016). *Citius, maius, melius*: del CREA al CORPES XXI. En J. Kabatek & C. de Benito Moreno (eds.), *Lingüística de corpus y lingüística histórica iberorrománica* (pp. 197-212). De Gruyter.

Rosenzweig, R. (2001). The Road to Xanadu: Public and Private Pathways on the History Web. *The Journal of American History*, *88*(2), 548-579.

Rovira Soler, J. C. (2001). Sobre los textos y sus derechos en la Biblioteca Virtual Miguel de Cervantes en la Universidad de Alicante. *Métodos de Información*, *8*(44), 67-70. https://doi.org/10.5557/%25x

Rovira Soler, J. C. y Rovira-Collado, J. (2019). La literatura en español en Internet: veinte años de la creación de la Biblioteca Virtual Miguel de Cervantes- *Mi Biblioteca*, *58*, 54-59.

Ruiz García, E. (2002). *Introducción a la codicología*. Fundación Germán Sánchez Ruipérez.

Russell, S. y Norvig, P. (2021). *Artificial Intelligence: a Modern Approach*. Pearson.

Rydberg-Cox, J. A. (2009). Digitizing Latin Incunabula: Challenges, Methods, and Possibilities. *Digital Humanities Quartely*, *3*(1).

Sahami, M. *et al.* (1998). A Bayesian Approach to Filtering Junk E-Mail. En *AAAI'98 Workshop on Learning for Text Categorization* (pp. 55-62). AAAI.

Sale, A. E. (2000). Lorenz and Colossus. En *Proceedings of the 13th IEEE Computer Security Foundations Workshop* (pp. 216-222). IEEE. https://doi.org/10.1109/CSFW.2000.856938

Sánchez, J. A. *et al.* (2013). tranScriptorium: A European Project on Handwritten Text Recognition. En 13th ACM symposium on Document Engineering. AMC. https://doi.org/10.1145/2494266.2494294

Sánchez Herrador, M. A. *et al.* (2010). El deterioro del libro antiguo como fuente de información histórica. *Revista General de Información y Documentación*, *20*, 281-296.

Sánchez Sánchez, M. y Domínguez Cintas, C. (2007). El banco de datos de la RAE. CREA y CORDE. *Per Abbat: boletín filológico de actualización académica y didáctica*, *2*, 67-70.

Schantz, H. F. (1982). *The History of OCR: Optical Character Recognition*. Recognition Technologies Users Association.

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quartely*, *11*(2).

Schweikert, G. *et al.* (2009). mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, *19*(11), 2133-2143. https://doi.org/10.1101/gr.090597.108

Seidel, R. W. (2000). Reconstructions, historical and otherwise: the challenge of high-tech artifacts. En R. Rojas y U. Hashagen (eds.), *The First Computers: History and Architectures* (pp. 33-52). MIT Press.

Seuret, M. *et al.* (2019). Dataset of Pages from Early Printed Books with Multiple Font Groups [Ponencia]. En *Proceedings*

of the 5th International Workshop on Historical Document Imaging and Processing (pp. 1-6). ACM. https://doi.org/10.1145/3352631.3352640

Severin, D. (ed.) (1990). *El cancionero de Oñate-Castañeda*. Seminary of Hispanic Medieval Studies.

Severin, D. (ed.) (2000). *Two Spanish Songbooks: the Cancionero Capitular (SV2) and the Cancionero de Egerton (LB1)*. Liverpool University Press.

Severin, D. (2004). *Del manuscrito a la imprenta en la época de Isabel la Católica*. Reichenberger.

Severin, D. (2007). The Four Recensions of fray Íñigo de Mendoza's *Vita Christi* with Some Unpublished Stanzas. En A. Deyermond y B. Taylor (eds.), *From the Cancioneiro da Vaticana to the Cancionero general: Studies in Honour of Jane Whetnall* (pp. 225-234). Queen Mary University of London.

Severin, D. (2017). Fray Íñigo de Mendoza del manuscrito a la imprenta: ¿un caso de autocensura? En *Variación y testimonio único: la reescritura de la poesía* (pp. 291-303). Universidad de Alicante.

Shafait, F. *et al.* (2008a). Document cleanup using page frame detection. *International Journal of Document Analysis and Recognition*, *11*, 81-96. https://doi.org/10.1007/s10032-008-0071-7

Shafait, F. *et al.* (2008b). Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(6), 941-954. https://doi.org/10.1109/TPAMI.2007.70837

Shannon, C. E. y Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Sharma, D. *et al.* (2007). Image compression and feature extraction with neural network. En *Proceedings of the Academy of Information and Management Sciencies v. 11* (pp. 33-37). AAIC.

Sharma, S. *et al.* (2020). Face Recognition System Using Machine Learning Algorithm. En *5th International Conference on Communication and Electronics Systems* (pp. 1262-1168). IEEE. https://doi.org/10.1109/10.1109/ICCES48766.2020.9137850

Snowden, R. *et al.* (2014). *Basic Vision. An Introduction to Visual Perception*. Oxford University Press.

Soler Fuensanta, J. R. (2004). Mechanical Cipher Systems in the Spanish Civil War. *Cryptologia*, *28*(3), 265-276.

Spiegel, J. *et al.* (2000). The ENIAC: history, operation, and reconstruction in VLSI. En R. Rojas y U. Hashagen (eds.), *The First Computers: History and Architectures* (pp. 107-120). The MIT Press.

Springmann, U. y Lüdeling, A. (2017). OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, *11*(2).

Stamatatos, E. (2008). A survey of modern authorship attribution methods. *JASIST*, *60*(3), 538-556.

Stokes, P. A. *et al.* (2017). The eScriptorium VRE for Manuscript Cultures. *Classical Journal*, *11*(2).

Szeliski, R. (2011). *Computer Vision*. Springer.

Tacón Clavaín, J. (2008). *La conservación en archivos y bibliotecas*. Ollero & Ramos.

Tolba, A. S. *et al.* (2006). Face Recognition: A Literature Review. *International Journal of Signal Processing*, *2*(2), 88-103.

Torruella, J. (1992). *La rima en la lírica medieval (estudi mètric del cançoner L)*. Universitat Autònoma de Barcelona.

Torruella, J. (2003). Arxiu informatitzat de textos catalans medievals. En *Actes del novè Col·loqui Internacional de Llengua i Literatura Catalanes vol. 2* (pp. 239-252). Abadia de Montserrat.

Torruella, J. y Lawrance, J. N. H. (1990). Un projecte d'arxiu informatitzat de textos catalans medievals: algunes normes. *Llengua i Literatura*, *3*, 481-506.

Turing, A. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. En *Proceedings of the London Mathematical Society vol. s2-42* (pp. 230-265). London Mathematical Society. https://doi.org/10.1112/plms/s2-42.1.230

Typenrepertorium der Wiegendrucke [TW]. (4 de mayo de 2023). Staatsbibliothek zu Berlin. https://tw.staatsbibliothek-berlin.de

Universal Short Title Catalogue [USTC]. (29 de abril de 2023). University of St Andrews. https://www.ustc.ac.uk

Valls i Subirà, O. (1978). *La historia del papel en España*. Empresa Nacional de Celulosas.

Vanhoutte, E. (2013). The Gates of Hell: History and Definition of Digital | Humanities |Computing. En M. Terras *et al.* (eds.), *Defining Digital Humanities: A Reader* (pp. 119-156). New York Routledge.

Verdejo, M. F. (1976). Un robot capaz de dialogar en castellano. *Boletín del Centro de Cálculo de la Universidad Complutense*, *29*, 1-16.

Vergara, J. (2002). *Conservación y restauración de material cultural en archivos y bibliotecas*. Biblioteca Valenciana.

Vindel, F. (1945-1954). *El arte tipográfico en España durante el siglo* XV. Dirección General de Relaciones Culturales.

Weaver, W. (1955). Translations. En W. N. Locke y A. D. Booth (eds.), *Machine translation of languages* (pp. 15-23). John Wiley & Sons.

Weichselbaumer, N. (2020). New Approaches to OCR for Early Printed Books. *Digitalia*, *2*, 74-87. https://doi.org/10.36181/digitalia-00015

Whinnom, K. (1961). Ms. Escurialense K-III-7: el llamado *Cancionero de Fray Íñigo de Mendoza*. *Filología*, *7*, 161-172.

Whinnom, K. (1962). The Printed Editions and Text of the Works of Fray Íñigo de Mendoza. *Bulletin of Hispanic Studies*, *39*, 137-152.

Whinnom, K. (1977). Fray Íñigo de Mendoza, fra Jacobo Maza, and the affiliation of some early Mss of the *Vita Christi*. *Annali di Ca' Foscari*, *16*, 129-139.

Wick, C. *et al.* (2020). A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. *Digital Humanities Quarterly*, *14*(2).

Williams, M. R. (1985). *A History of Computing Technology*. Prentice-Hall.

Wisbey, R. (1965). Computers and Lexicography. En D. H. Hymes (ed.), *The use of computers in anthropology* (pp. 215-234). De Gruyter Mouton. https://doi.org/10.1515/9783111718101.215

Zarco Cuervas, J. (1926). *Catálogo de los manuscritos castellanos de la Real Biblioteca de El Escorial*. Imprenta Helénica.

Zerdoun Bat-Yehouda, M. (1983). *Les encres noires au Moyen âge (jusqu'à 1600)*. Centre National de la Recherche Scientifique.

Zuse, H. (2013). Reconstruction of Konrad Zuse's Z3. En A. Tatnall, T. Blyth y R. Johnson (eds.), *Making the History of Computing Relevant* (pp. 287-296). Springer. https://doi.org/10.1007/978-3-642-41650-7_26


ULTREIA
EDITORIAL