ESTADO DE LA DIGITALIZACIÓN DE *LA* EDAD DE PLATA: UN ANÁLISIS CUANTITATIVO

STATE OF THE DIGITIZATION OF THE EDAD DE PLATA: A QUANTITATIVE APPROACH

José Calvo Tello
University of Würzburg
jose.calvo@uni-wuerzburg.de

Resumen

En este artículo analizo el panorama que el investigador interesado en trabajar con textos digitales de la Edad de Plata encuentra al comenzar su trabajo. He realizado búsquedas sistemáticas de autores en siete fuentes generalistas de textos y documentos digitales: Project Gutenberg, Biblioteca Digital Hispánica de la BNE, Biblioteca Virtual Cervantes, Internet Archive, Google Books, Wikisource y EpubLibre. La nómina de autores contiene aquellos que publicaron obras de prosa entre 1880 y 1939: un total de 135 que van (según su fecha de nacimiento) desde Valera hasta Francisco de Ayala. Los principales objetivos de este trabajo son: 1) valorar según diferentes criterios los proyectos de digitalización y publicación electrónica; 2) tener una imagen más ajustada del estado de la digitalización de esta época; 3) observar qué criterios ha seguido la digitalización de estos autores. Finalmente, trato de concretar las oportunidades y problemas que debemos tener en cuenta para el diseño y desarrollo futuro de nuestro trabajo.

Palabras clave: Digitalización. Distant Reading. Textos. Edad de Plata. XHTML.

Abstract

In this article, I analyze the situation that the researcher finds if she/he is interested in working with digital texts of the Edad de Plata. I have conducted systematic searches of authors in seven general repositories of texts and digital documents: Project Gutenberg, Biblioteca Digital Hispánica of the BNE, Biblioteca Virtual Cervantes, Internet Archive, Google Books, Wikisource and EpubLibre. The list of authors contains those who published prose works between 1880 and 1939: a total of 135 authors (sorted by their date of birth) from Valera to Francisco de Ayala. The main objectives of this work are: 1) to assess, according to different criteria, the digitization and electronic publishing projects; 2) to gain a more accurate picture of the state of digitization of this period; 3) to observe what criteria have conducted the digitization of these authors. Finally, I try to sum up the opportunities and problems that must be considered for the design and development of future work.

Keywords: Digitalization. Distant Reading. Texts. Edad de Plata. XHTML.

1. INTRODUCCIÓN

Digitalización, *big data* y *distant reading* concentran el interés de miles de proyectos institucionales y empresariales alrededor del mundo. El primero está relacionado con las iniciativas que bibliotecas y archivos han desarrollado en las últimas décadas por conservar y dar acceso en formato digital a sus fondos (Agenjo, 2015: 12-15), en su mayoría produciendo imágenes o archivos PDF. *Big data* está asociado con las metodologías de aprendizaje automático para fines como la investigación pública o privada, periodismo de datos, espionaje, etcétera (Schöch, 2013: 2-13). Las Humanidades Digitales han acuñado el concepto de *distant reading* (Moretti, 2005) que recoge una nueva manera de acercarse a los datos humanísticos en un volumen mucho mayor a lo que los enfoques tradicionales eran capaces. Estos tres conceptos, aunque diferentes, están relacionados: la digitalización es una manera de conseguir *big data* con la que aplicar metodologías de *distant reading.*

En los últimos años diferentes metodologías que utilizan volúmenes grandes de textos se han popularizado en los círculos internacionales de Humanidades Digitales. Algunos de ellos han sido desarrollados por investigadores de esta área, como Delta, uno de los principales algoritmos estilométricos, propuesto por Burrows (2002: 267-287). Este algoritmo ha sido implementado en un paquete de R llamado *stylo* (Eder *et al*., 2016: 1-15), explicado con ejemplos en español en Calvo Tello (2016: 140-176) y utilizado para analizar casos de atribución de autoría en textos españoles como el *Quijote de Avellaneda* (Rißler-Pipka, 2016:

27-51), el *Lazarillo* (de la Rosa y Suárez, 2016: 373-438) o *La conquista de Jerusalén* de Cervantes (Calvo Tello y Cerezo Soler, en prensa). Otros han sido adoptados de otras áreas de investigación, como Topic Modeling, explicado por Blei (2012: 77-84) y aplicado por Navarro-Colorado (2015a: 105-113) a un corpus de sonetos del Siglo de Oro; o grafos de relaciones de personajes (Trilcke *et al*., 2015).

Las instituciones de diferentes países, entre ellas las españolas, han invertido grandes esfuerzos en la digitalización como recoge Agenjo (2015: 13). Sin embargo, sus objetivos y formatos no son en muchos casos lo que los investigadores humanistas (más o menos digitales) necesitan (Lucía Megías, 2012; Allés Torrent, 2015: 18-21).

Encontrar imágenes o PDFs de cada página de un libro digitalizado queda muy lejos de poder aplicar a ese contenido técnicas de análisis textual: el investigador debe recolectar y ordenar las diferentes partes (si no lo estaban), aplicar algún programa de reconocimiento de caracteres (si no se ha hecho o si el resultado es deplorable), asegurarse de que el resultado es aceptable, limpiarlo de paratextos innecesarios para el análisis y finalmente convertirlo en texto plano o (si el análisis lo necesita) en XML-TEI. Este es el trabajo que el investigador tiene por delante si comienza su trabajo con digitalizaciones provenientes de la BNE, Internet Archive o Google Books (y numerosos textos de la Biblioteca Cervantes Virtual, como, por ejemplo, la mayoría de las novelas de Blasco Ibáñez, por señalar un ejemplo).

En paralelo a estas iniciativas se han desarrollado otros proyectos que han codificado textos literarios en lenguajes de marcado, principalmente (X)HTML, entre los que cabe mencionar la Biblioteca Cervantes Virtual, Wikisource o ePubLibre.

Un tercer grupo de proyectos ha codificado y publicado textos literarios de un género, época o autor concreto, o en una cantidad reducida. Algunos de ellos han publicado sus resultados en (X)HTML (en algunas ocasiones en formato de libro electrónico ePUB): *Teatro Español del Siglo de Oro* (1998), *An Electronic Corpus of 15th Century Castilian Cancionero Manuscripts* (Severin *et al*., 2007), la colección *Clásicos Hispánicos* (Jauralde Pou *et al*., 2012) o *Biblioteca Digital Artelope* (2013). La minoría han dado acceso abierto a sus versiones de XML-TEI, como *IMPACT-es Diachronic Corpus* (Sánchez-Martínez, *et al*. 2013), *Moralische Wochenschriften* (Semlak, 2014), *Corpus of Spanish Golden-Age Sonnets* (Navarro-Colorado, 2015b) o los corpus de novelas de *CLiGS Textbox* (Schöch, 2015). Recientemente, el grupo de investigación GHEDI ha anunciado la creación de una *Biblioteca electrónica textual del teatro en español* (Gómez *et al*., 2015: 171-184) también en formato XML-TEI y con textos de la Edad de Plata.

Como se observa, aunque se pueden listar numerosas iniciativas que dan acceso a los documentos textuales o a los mismos textos en formato digital, falta un proyecto generalista que dé acceso a los textos en el formato de codificación filológica digital estándar que es XML-TEI. Nuestros colegas que están trabajando en textos literarios ingleses o

alemanes pueden acceder a repositorios como Oxford Text Archive o TextGrid donde pueden encontrar miles de textos en formato XML-TEI, a texto completo y bajo licencias Creative Commons. Tener acceso a este tipo de recursos permite poder aplicar metodologías de *distant reading* como las arriba mencionadas. Hasta que no haya un proyecto que aborde esta tarea (ya sea que el Cervantes Virtual deje de negar el acceso al formato XML-TEI, ya sea que otra iniciativa asuma la tarea) nos veremos obligados a asumir el trabajo de crear nuestro propio corpus, o desistiremos de aplicar *distant reading*, que es la situación actual de numerosos proyectos.

En el grupo de investigación CLiGS en la Universidad de Würzburg decidimos construir nuestros propios corpus o colecciones de textos en formato XML-TEI a partir de las diferentes ediciones electrónicas que encontramos o digitalizamos nosotros mismos. Las herramientas que hemos utilizado para la conversión de diferentes formatos de HTML a XML-TEI están publicadas en nuestro repositorio de GitHub Toolbox (Schöch *et al*., 2014). Parte del fruto de ese trabajo ya ha sido publicado (formatos XML-TEI y texto plano) en el repositorio *Textbox* (Schöch, 2015), que contiene cuatro colecciones de textos: dos de ellas en francés (novelas y novelas breves) y dos en español: uno sobre novelas latinoamericanas (preparado por Ulrike Henny) y uno de novelas españolas (preparado por Calvo Tello).

2. METODOLOGÍA: NÓMINA DE AUTORES Y DE REPOSITORIOS

Una vez trazadas las principales iniciativas y problemas de los textos electrónicos en español, voy a describir la metodología seguida para analizar el estado de digitalización de la Edad de Plata. Esta queda también documentada en la hoja de cálculo que forma el apéndice de este trabajo y que contiene todos los datos (incluidas las URIs de cada autor en cada repositorio) de los que este artículo parte.

Frente al consenso de que la Edad de Plata se termina con la Guerra Civil (cabe discutir si con el comienzo o el final), su inicio es menos claro. Hay dos fechas principales que se utilizan para delimitar la Edad de Plata: 1902 (Mainer, 1981) y 1868 (Urrutia Cárdenas, 1999: 581-595). La primera deja fuera algunas de las primeras obras de autores del 98 como Unamuno o Valle, mientras que la segunda prácticamente dobla la extensión de este período. Es por eso que decidí llegar a un término intermedio extenso que englobase también la mayor parte de la creación de algunos de los principales autores del período realista-naturalista como son Galdós, Bazán o Clarín. La fecha elegida fue el año 1880.

Debido a que mi tesis doctoral se centra en el estudio de los subgéneros de prosa, en la nómina de autores analizados se encuentran solo aquellos que escribieron obras de prosa entre los años 1880 y 1939. Para saber qué autores escribieron obras en esta época, utilicé la colección de volúmenes de *Manual de Literatura Española* de Pedraza Jiménez y Rodríguez

Cáceres desde el séptimo hasta el décimo tercero[1]. Solamente recogí aquellos autores de los que los manuales señalan de su obra más información que meramente los títulos. En concreto, para incluir al autor en la nómina, los manuales debían señalar que alguna de sus obras pertenecía a algún género de prosa. Aunque esto pueda parecer una fuerte limitación para el estudio, se utilizó un concepto muy amplio de prosa aceptando textos parcialmente en prosa (*Diario de un poeta recién casado*, *Azul.*) o subgéneros de prosa poco prototípicos como el cuento, la greguería, la memoria, la biografía o el libro de viajes. Por lo tanto, se encuentran muchos autores que no asociamos como prosistas, como Azaña, Cernuda, Darío, d'Ors, Juan Ramón Jiménez, Lorca, los hermanos Machado, Maeztu, Miguel Hernández, Panero, Poncela, Rosalía de Castro o Salinas. En total la nómina incluye 135 autores.

Todos estos autores fueron sistemáticamente buscados en diferentes proyectos de digitalización de libros o publicación de textos. Los siete portales o repositorios[2] que han sido investigados son: Project Gutenberg, Biblioteca Digital Hispánica de la BNE[3], Biblioteca Virtual Cervantes, Internet Archive, Google Books, Wikisource y EpubLibre.

La gran mayoría de estos repositorios son suficientemente conocidos por la comunidad de Humanidades Digitales. Ya hemos mencionado más arriba los proyectos de digitalización de libros. Debido al carácter diferente de cada uno de estos repositorios y sus posibilidades de búsqueda, en algunos casos, las búsquedas son más específicas que en otros. Por ejemplo, mientras que en Internet Archive he realizado la búsqueda sin restricciones, en Google Books he buscado exclusivamente por libros del autor que estén a texto completo mediante el filtro *eBooks gratuitos*. Para obtener más detalle sobre las búsquedas en cada repositorio y cada autor, pueden observarse los parámetros de las URIs en el documento anexo de este artículo.

En cuanto a los proyectos de publicación de textos, el más importante por su carácter pionero (25 años anterior al siguiente de los aquí mencionados) e internacional es el Project Gutenberg. En el ámbito hispánico es de notable importancia la Biblioteca Cervantes Virtual, proyecto que ha codificado miles de textos en XML-TEI pero que no da acceso a ellos, sino exclusivamente a versiones en HTML. Otros proyectos notables son Wikisource (repositorio emparentado con Wikipedia, que almacena textos) o ePubLibre.

Probablemente este último sea el proyecto menos conocido, por lo que creo necesario aportar algunos datos sobre él. Es un proyecto de publicación de libros electrónicos que hasta la fecha ha publicado unos 26.000 eBooks, principalmente en español. La información sobre personas o instituciones responsables es escasa en su página web, probablemente debido a

[1] Este último volumen tiene el foco en la literatura de posguerra, pero recoge obras publicadas antes de la guerra.
[2] Aunque el concepto de repositorio puede ser más limitado, voy a preferir este concepto en el resto del artículo para hacer referencia a cualquiera de estos portales.
[3] A partir de ahora simplemente mencionada como BNE.

los posibles problemas legales por infracción de propiedad intelectual. Los eBooks de este portal se pueden descargar mediante *torrents* (tecnología P2P). El proyecto está activo desde 2013, donde recogió el testigo de la web epubgratis.me (desaparecida por completo en 2014). Han publicado un *Manifiesto*[4] en el que explican lo siguiente:

> *Este proyecto nace del deseo de lectores anónimos de compartir sus libros favoritos con todo aquel que sienta ansias de lectura, ya que la cultura debe ser un bien universal accesible a todo el mundo. […] Los editores realizamos todo este trabajo sin pedir nada a cambio, y sin más deseos de recompensa que tu disfrute. Jamás con ánimo de lucro.*

Para saber más sobre el proyecto probablemente sea útil leer los posts del blog *Más allá de las Puertas de Tannhäuser*[5], aunque es difícil certificar la veracidad de los datos en un proyecto tan anonimizado.

Para realizar el análisis, el repositorio debía haber publicado al menos un texto (no necesariamente de prosa) completo[6] del escritor como principal autor[7]. Esto quiere decir que este análisis simplifica los datos para señalar no cuántos textos contiene de cada autor o cuál es su calidad, sino simplemente si el repositorio contiene al menos un texto completo o no. Las búsquedas se realizaron entre los meses de enero y junio de 2016. Como ya he señalado más arriba, todos los datos de los que parte este artículo se encuentran en la hoja de cálculo que forma el apéndice de este trabajo.

3. ESTADO Y VALORACIÓN DE LOS REPOSITORIOS

Una vez busqué todos los autores en todos los repositorios, realicé algunos cálculos de estadística descriptiva. En primer lugar, veamos los valores según proyectos de digitalización:

[4] Accesible desde https://epublibre.org/inicio/manifiesto.
[5] Accesible desde https://rosmar71.wordpress.com/?s=epub+libreandx=0andy=0.
[6] Es decir, no era suficiente que el autor fuese coautor de algún texto como antologías de cuentos, caso muy frecuente en Project Gutenberg.
[7] Algunos proyectos como Cervantes Virtual o ePubLibre contienen información sobre si la persona fue traductor, prologuista o destinatario del texto.

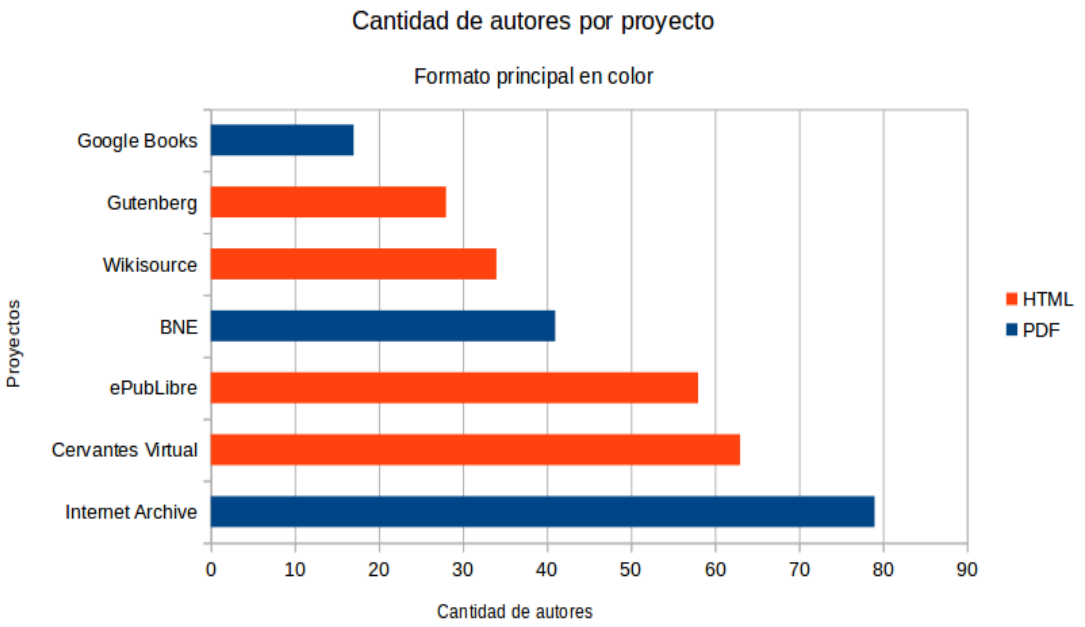

Figura 1. Cantidad de autores por repositorio.

Como se puede observar, Internet Archive es el repositorio que más autores diferentes contiene, con un total de 79 autores. Los dos siguientes proyectos se llevan poca distancia entre sí, Cervantes Virtual (63) y ePubLibre (58). El paso al siguiente proyecto desciende en unos 20 autores, con lo que encontramos la BNE (41), seguida de Wikisource (28) y Gutenberg (28). De manera sorprendente nos encontramos a Google Books cerrando la clasificación. Asimismo, llama la atención el hecho de que el tipo de formato (PDF o HTML) no parece definir el volumen de los autores publicados. Debido a que producir imágenes requiere menos trabajo de edición y limpieza que publicar lenguaje de marcado, se hubiese podido esperar que estos proyectos copasen los primeros puestos.

En segundo lugar, me he preguntado no por la cantidad sino por la innovación que cada proyecto representa respecto a los demás. Es decir, ¿cuántos autores puedo encontrar en cada repositorio que no puedo encontrar en los otros seis portales?:

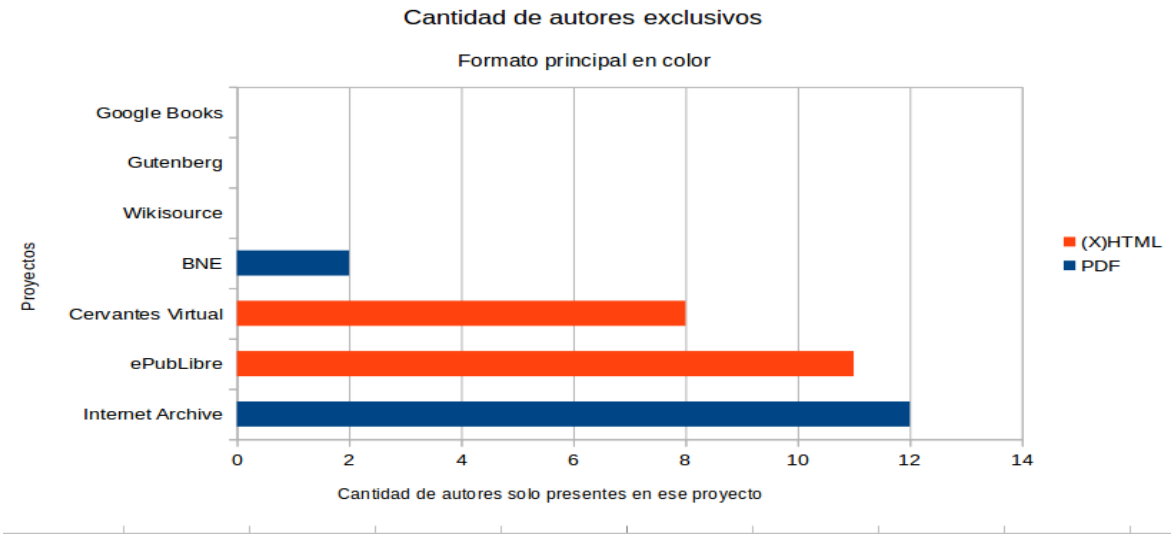

Figura 2. Cantidad de autores exclusivos en cada repositorio.

En este caso se ven claramente tres grupos diferentes: Internet Archive, ePubLibre y Cervantes Virtual (algo más separado) rondan los 10 autores que ofrecen de manera exclusiva frente al resto de repositorios. El portal digital de la BNE contiene dos únicos autores no contenidos por el resto. Por último, los otros tres proyectos no contienen ningún autor *en exclusiva*.

Si comparamos las dos gráficas notaremos dos aspectos llamativos. En primer lugar, el orden de los proyectos apenas varía: solamente ePubLibre y Cervantes Virtual pivotan en sus posiciones entre el segundo y tercer puesto. Internet Archive queda en ambas como el primer proyecto, la BNE queda en cuarto puesto y Wikisource, Gutenberg y Google Books copan los últimos puestos. En segundo lugar, es sorprendente que la primera gráfica, que representa una transición moderada, pase a convertirse a una gráfica tan abrupta como la segunda.

La interpretación de los datos no parece sencilla y puede deberse a diferentes causas. Es posible que Google Books contenga muchos más autores de los que se recoge aquí pero que por estrategia empresarial (publicación por Hathi Trust)[8] o problemas legales no los muestre a través de su portal de libros. Wikisource nunca ha sido el proyecto bandera de Wikimedia, por lo que es posible que su labor se haya limitado a absorber los autores que ya habían sido publicados por otros proyectos. Gutenberg en cambio, es el proyecto más antiguo y conocido de los aquí analizados, por lo que puede ser que sus textos hayan sido absorbidos por el resto de proyectos o que su nómina de autores haya servido de modelo para otros. De cualquier manera, hay que recordar que la unidad analizada es autor y no texto, por lo que sí

[8] El objetivo de este trabajo es analizar qué proyectos abiertos puede utilizar el investigador, por lo que Hathi Trust queda descartado.

puede ocurrir que encontremos textos concretos en Google Books, Gutenberg o Wikisource que no encontraremos en el resto de los proyectos.

4. ESTADO DE LA DIGITALIZACIÓN DE LA EDAD DE PLATA

En esta sección quiero analizar el estado de la digitalización de la Edad de Plata según la nómina de autores arriba mencionada. Mi objetivo en este punto ya no es aportar datos sobre cada repositorio en concreto, sino describir el estado que ya se ha realizado y el que queda por hacer. Para ello he utilizado principalmente conceptos básicos de estadística descriptiva como tendencias de centralidad e histogramas.

En primer lugar, vamos a conseguir los valores de tendencia de centralidad sobre la cantidad de proyectos en los que aparece un autor:

- media: 2,37
- mediana: 2
- moda: 1
- desviación típica: 2,05

El concepto más familiar entre las tendencias de centralidad es el de media, en este caso de 2,37. Es decir, la nómina de autores tiene una media de 2,37 repositorios en los que los autores aparecen digitalizados. Si nos quedáramos con este número tenderíamos a pensar que la digitalización ha sido completada. Sin embargo, sabemos que la media es sensible a valores atípicamente altos o bajos en caso de que los haya. Para tales casos la mediana es más robusta, valor que en este caso aporta un valor de 2. Aunque el valor se ha reducido, vemos que sigue siendo positivo. La moda, o el valor más frecuente, es de 1. Es decir, lo que más frecuentemente ocurre en concreto es que un autor se encuentre digitalizado por un único proyecto. Finalmente nos interesa observar cuál es la desviación típica, valor que recoge si la distribución de los datos está concentrada cerca de la media o si por lo contrario hay numerosos valores atípicamente altos o bajos. Su valor es de 2,05, lo que debe interpretarse como que los valores 2,05 por debajo o por encima de la media no son excepcionales. Es decir, no es excepcional que un autor se encuentre digitalizado por 0 repositorios (0,32 en concreto) o por 4 (4,42 en concreto).

Para tener una idea más global, observemos el histograma de la distribución de la cantidad de autores en proyectos de digitalización:

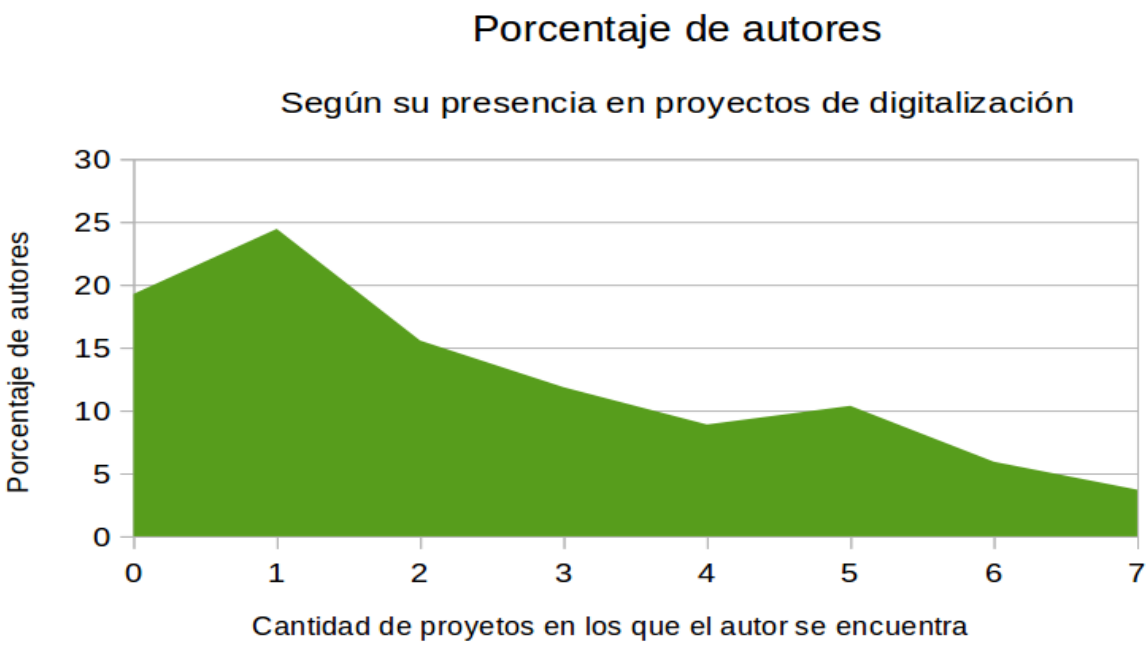

Figura 3. Histograma de porcentajes de autores en cantidad de repositorios.

Como vemos, la forma del histograma es asimétrica a la izquierda y es bimodal al tener dos picos en 1 y 5. Ahora sabemos que casi el 20% de los autores no han sido digitalizados por ningún proyecto y que menos del 5% aparecen en todos los proyectos de digitalización analizados. Si seguimos la distribución del histograma, observamos una bajada abrupta (casi un 9%) entre uno y dos repositorios. Si aceptamos la idea intuitiva de que muchos de estos proyectos no vuelven a digitalizar los textos, sino que reutilizan el trabajo de otros repositorios, diremos que no ocurre de manera inmediata que el segundo proyecto absorba los nuevos autores; una vez está en el segundo proyecto, otros proyectos tienden a absorberlo más rápidamente, hasta llegar a 5 proyectos, donde encontramos un pico en el histograma. En ese punto volvemos a ver una bajada brusca. Recordemos de cualquier manera que la media era de 2,4 y la desviación estándar superior a 2, por lo que en realidad esa moda secundaria de estabilización entre proyectos digitalizadores es un valor excepcional. Una vez que tenemos una imagen general sobre la digitalización, observamos los mismos valores, pero diferenciando los proyectos entre los de digitalización que producen PDFs (Internet Archive, BNE y Google Books) y los de publicación de textos electrónicos en lenguaje de marcado (X)HTML (Gutenberg, Wikisource, ePubLibre y Cervantes Virtual)[9]. En PDF:

[9] Como ya he señalado, Cervantes Virtual ofrece, en realidad, PDFs, textos en HTML o incluso enlaces a fuentes externas, por lo que es una mezcla entre catálogo, proyecto de digitalización y proyecto de publicación de textos electrónicos. Sin embargo, para los textos literarios, la mayoría de ellos se encuentran en HTML.

- media: 1,01
- mediana: 1
- moda: 0
- desviación típica: 0,98

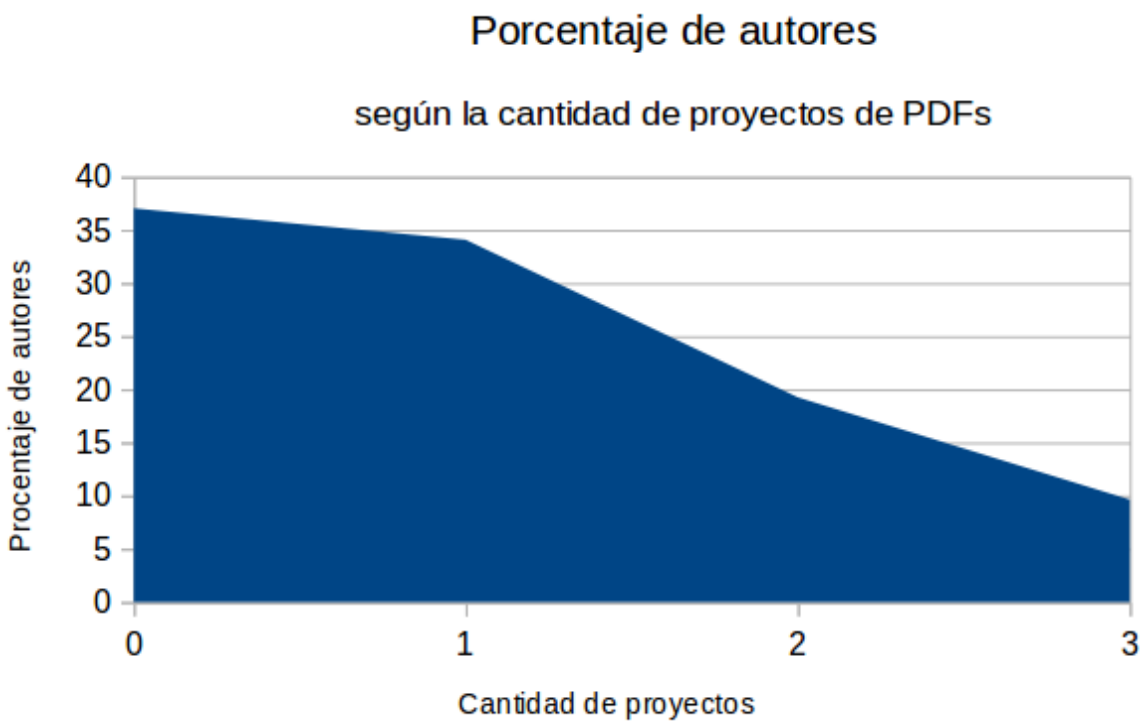

Figura 4. Histograma de porcentajes de autores en cantidad de repositorios de PDFs.

En (X)HTML:

- Media: 1,36
- Mediana: 1
- Moda: 0
- Desviación típica: 1,31

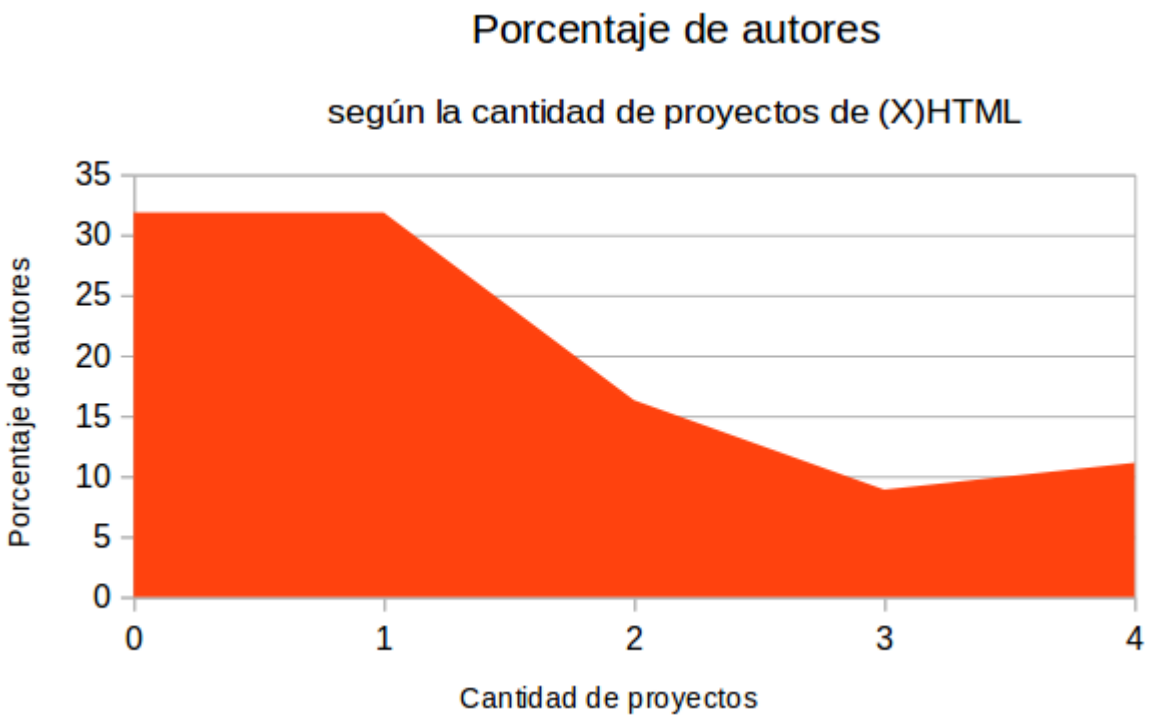

Figura 5. Histograma de porcentajes de autores en cantidad de repositorios de (X)HTML.

Observando ambos conjuntos de datos e histogramas, percibimos que, para un formato concreto, la cantidad de autores presentes en ningún proyecto aumenta hasta convertirse para ambos formatos en modas, aunque 1 está a muy poca distancia en el caso de PDFs y tiene el mismo valor que 0 en (X)HTML. Mientras que el desarrollo del histograma de PDFs desciende de manera más moderada, el de (X)HTML desciende de manera abrupta hasta 2 obteniendo una moda menor en 4. Es decir, los recursos de otros proyectos de publicación de texto en lenguajes de marcado podrían reutilizarse más para aumentar su propio catálogo.

5. TENDENCIAS EN LA DIGITALIZACIÓN: ÉPOCA Y CANON

Esta última sección sobre datos e interpretación se centra en observar qué tendencias se han seguido de manera generalizada en la digitalización y publicación de textos de estos autores. He intentado analizar dos hipótesis diferentes:

1) Muchos de los autores que escribieron en la Edad de Plata aún no están en dominio público. ¿Afecta la fecha de su muerte en su estado de digitalización? En concreto ¿afecta el hecho de que el autor se encuentre en dominio público?

2) ¿Son los autores más importantes (o los más canonizados) más digitalizados que los menos importantes?

Para analizar ambas hipótesis he visualizado los datos en un diagrama de dispersión, con el fin de saber si se observa una correlación lineal entre ambos conjuntos de valores. Después de comprobarlo, he calculado para ambos el coeficiente de correlación de Pearson (Haslwanter, 2016: 184).

Comencemos por la primera hipótesis, si el año de la muerte del autor tiene correlación con su estado de digitalización. En el siguiente diagrama de dispersión cada autor está representado por una burbuja. El eje horizontal representa el año de su muerte, mientras que el eje vertical refleja la cantidad de proyectos digitalizadores en los que se encuentra. El color de la burbuja representa si el autor se encuentra en dominio público en 2016 o no[10]:

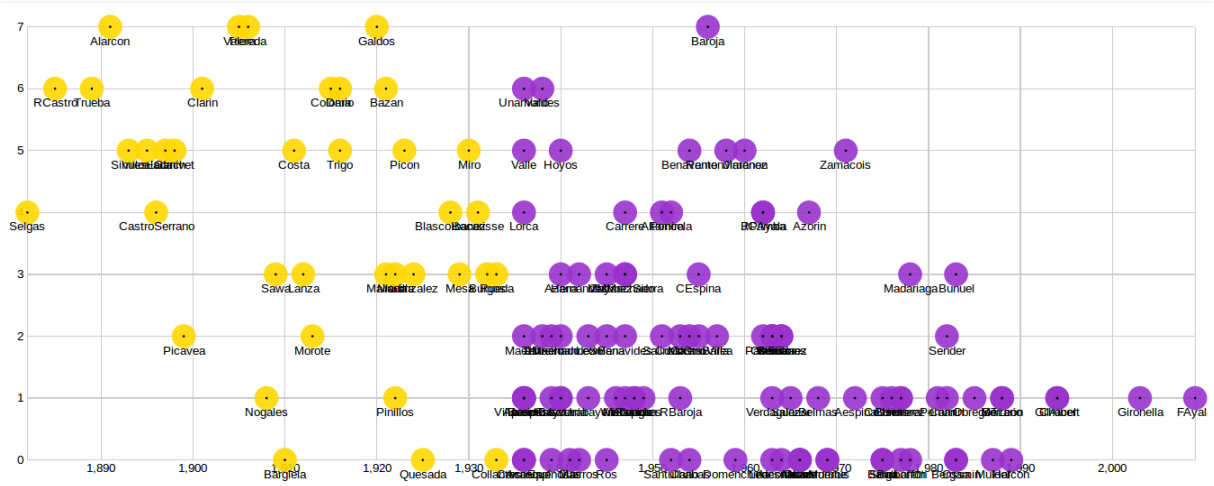

Figura 6. Diagrama de dispersión de la fecha de la muerte y la digitalización.

Como se puede observar, se aprecia cierta linealidad descendente en el diagrama al estar en más proyectos de digitalización aquellos autores fallecidos en el cambio del siglo XIX al XX, frente a los autores muertos tras la Guerra Civil, donde la digitalización se hace cada vez más infrecuente. Por supuesto hay numerosos extremos, tanto en la esquina inferior izquierda (Selgas, Picavea, Nogales, Bargiela) como en la esquina superior derecha (Unamuno, Valdés, Baroja, Zamocois). De cualquier manera, no se observa una correlación curvilínea que negaría la aceptabilidad del coeficiente de correlación de Pearson. Para observar cuán fuerte es esta correlación, he realizado el test de coeficiente de correlación de Pearson, que aporta un valor en -1 (perfecta correlación negativa) y 1 (perfecta correlación positiva), significando 0 que no hay ninguna correlación. Para este caso, el coeficiente aporta un resultado de -0,55. Esto quiere decir, según la guía sugerida por James (1996) que el año de la muerte del autor y su presencia en proyectos de digitalización tienen una correlación moderada negativa: cuanto más temprana sea la fecha de la muerte, en más repositorios de digitalización se encontrará. Para el caso concreto del dominio público, el test aporta un valor

[10] Los diagramas de dispersión han sido realizados en *Raw*, http://rawgraphs.io/.

de 0,51, es decir una correlación moderada positiva. Si el autor está en dominio público tiene más posibilidades de haber sido digitalizado.

Estos resultados no son sorprendentes: es la impresión intuitiva si hemos hecho búsquedas de autores muertos durante el siglo XX. Este artículo consigue constatar este hecho para estos autores. Si aceptamos que estos resultados son extrapolables a una franja temporal más amplia, otros países o lenguas, arrojaría importantes malas noticias para la investigación que quiera utilizar metodologías de *distant reading*, especialmente la interesada en la segunda mitad del siglo XX: durante los siguientes años tenderemos a no encontrar en Internet textos de autores que murieron tras la década de los años 30. Mediante este análisis cuantitativo demostramos para autores españoles de esta época la opinión que Jockers (2013: 175) expresa de manera general para la literatura: "scholars wishing to study the literary record at scale are forced to ignore almost everything that has been published since 1923". Estos resultados también parecen señalar el desinterés que tienen las fundaciones y herederos de los autores en que sus textos heredados estén accesibles en formato digital en estos repositorios.

Aun así, las correlaciones solo son moderadas. Sorprende que autores como Valle, Valdés, Baroja o Benavente estén presentes entre 5 y 7 repositorios a pesar de que sus herederos aún mantienen los derechos de explotación. De cualquier manera, hay que recordar de nuevo que este análisis no cubre la cantidad o calidad de las digitalizaciones, ni siquiera el tipo de obras digitalizadas. Muchos de los textos de esos autores presentes en BNE o Cervantes Virtual son en realidad textos breves, parciales o cartas. De la misma manera, hay autores fallecidos hace más de un siglo que apenas han sido digitalizados (como Picavea, Morote, Nogales, Bargiela, Pinillos o Quesada).

Pasemos ahora a la segunda hipótesis, si la importancia del autor mantiene correlación con su estado de digitalización. Para ello era necesario expresar el canon de una manera numérica. Por supuesto cualquier aproximación a enumerar el canon puede ser objeto de fuertes críticas. Para este análisis he decidido utilizar la cantidad de páginas que el *Manual de Literatura Española* dedica de manera específica a cada autor[11]. Este valor representa el eje horizontal en el siguiente diagrama de dispersión, siendo el resto de parámetros los mismos del anterior gráfico:

[11]Barajé otras maneras de contabilizar el canon como la cantidad de veces que el nombre del autor aparecería en N-Grams de Google o en corpus históricos como el CORDE. Sin embargo esto trae enormes problemas de desambiguación, además de que, principalmente en corpus equilibrados, la enorme mayoría de autores no aparecían ninguna vez, por lo que se perdía granularidad en el canon. Además de ser fácilmente accesible mediante el índice, la cantidad de páginas en el *Manual* tiene la ventaja de mantener coherencia con la formación de la nómina de autores.

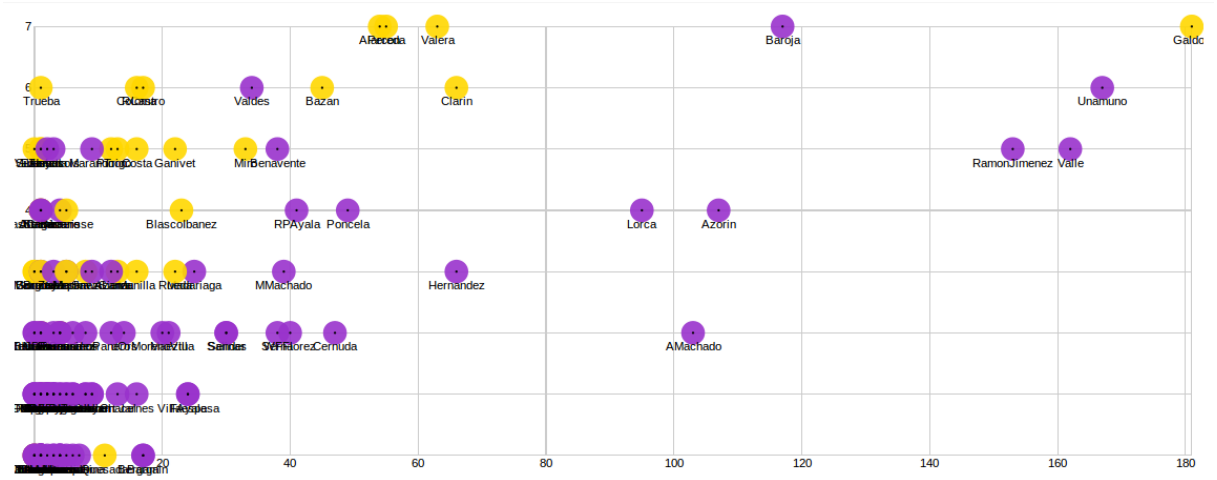

Figura 7. Diagrama de dispersión de canon y la digitalización.

Como observamos, la relación vuelve a ser lineal, con una gran variedad en aquellos autores menos canonizados. De nuevo vemos algunas excepciones: algunos autores menos canonizados como Trueba o Coloma disfrutan de presencia en numerosos repositorios, frente a otros autores que a pesar de estar canonizados apenas aparecen en algunos proyectos de digitalización, como Antonio Machado, Miguel Hernández, Cernuda, Serna o Fernández Flórez. Es interesante señalar que podemos explicar mediante ambos diagramas casos que en uno de ellos parecían excepciones: los autores canónicos menos digitales murieron tras el comienzo de la Guerra Civil; sin embargo, aunque Unamuno, Valle o Baroja murieron tras 1936, son demasiado canónicos como para no estar digitalizados. De todos modos, sigue habiendo casos a los que estas gráficas no parecen dar explicación, como la excepcional presencia de Valdés o Zamacois en numerosos repositorios.

Ahora que sabemos que la distribución es lineal, calculemos para estos datos el coeficiente de correlación de Pearson, que aporta un valor de 0,53; es decir, que esta formalización del canon y el estado de la digitalización mantienen una moderada correlación positiva: cuanto más canonizado, más digitalizado.

Si consideramos plausible que la digitalización de textos de otras épocas ha seguido parámetros similares, entonces concluiremos que en general es más fácil encontrar textos en Internet de aquellos autores que son considerados como los más importantes. Por supuesto en este artículo estoy probando dos hipótesis sobre las estrategias que ha seguido en general la digitalización de estos autores, pero eso no significa que la digitalización no haya seguido otros parámetros que no he probado. Si nos fijamos en la última gráfica, observaremos que de manera general los autores que escribieron principalmente prosa (Galdós, Unamuno, Baroja, Clarín, Bazán, Valera, Alarcón, Pereda, Valdés) están digitalizados en más proyectos que poetas y dramaturgos (claramente en Antonio Machado, Miguel Hernández, Cernuda; menos claro en Valle, posiblemente porque ejerció maestría en los tres géneros). Es decir,

parece que estas fuentes han preferido digitalizar a prosistas, pero este artículo no da cuenta exacta de estos resultados al necesitar datos cuantitativos sobre el género de cada obra de cada autor.

6. CONCLUSIÓN

Por último, quiero recopilar las ideas principales que este artículo presenta sobre el estado de digitalización de la Edad de Plata (1880-1939), utilizando como principal unidad de trabajo los autores (y no cada uno de los textos) y centrándonos en aquellos autores que publicaron obras en prosa y fueron recogidos por el *Manual de Literatura Española*[12]. Todos los datos de los que parte este artículo se encuentran en una hoja de cálculo que forma el apéndice de este.

En primer lugar, debemos recordar que la necesidad de un artículo para saber el estado de la digitalización en formato como (X)HTML y PDF se debe a que la comunidad investigadora de Humanidades Digitales hispánicas sigue sin un repositorio generalista que ofrezca textos literarios en un formato adecuado para la investigación. Esto debe ser entendido como un lastre para nuestra área que nos pone en desventaja con nuestros colegas anglistas o germanistas. En los últimos años sin embargo se han publicado colecciones y corpus de textos literarios codificados en XML-TEI, manera en la que se debería continuar trabajando.

En cuanto a los siete proyectos de digitalización y publicación de textos electrónicos, las principales conclusiones son:

- Cuantitativamente, Cervantes Virtual, ePubLibre e Internet Archive son los repositorios más amplios y con más autores que no encontraremos en otros proyectos. Lamentablemente los dos últimos son menos conocidos por la comunidad investigadora.
- BNE resulta el escalón intermedio entre los tres grupos de repositorios según los dos criterios analizados.
- Gutenberg, Wikisource y Google Libros, tres repositorios muy conocidos por la comunidad, no se destacan ni en cuanto a la cantidad de autores ni aportan autores que no encontremos en los otros proyectos.

[12] Es decir, este trabajo no pretende ser una aproximación al estado de la digitalización de la literatura española en general, aunque algunas de sus conclusiones podrían servir como orientación parcial para otros ámbitos.

En relación al estado de la digitalización en PDF y de publicación en (X)HTML, hemos observado que:

- Alrededor del 80% de los autores de la nómina tiene al menos un texto en uno de los repositorios.
- Los autores que aparecen en cinco o más de estos repositorios son excepcionales.
- No es excepcional que un autor aparezca entre ningún y cuatro repositorios.
- Para el formato (X)HTML (preferible para el investigador al no disponer de XML-TEI), la situación empeora, siendo lo más habitual encontrar ninguno o un proyecto que haya digitalizado un autor concreto.
- Los proyectos de publicación de textos podrían reutilizar más los fondos de otros repositorios.
- El último punto de análisis del artículo trataba de observar qué tendencias ha seguido la digitalización y publicación de textos.
- La digitalización de estos autores está en correlación moderada con el año de su muerte: cuanto más tiempo hace que el autor murió, más digital; el dominio público también está en correlación con la digitalización.
- La digitalización de estos autores está en correlación moderada con el canon: cuanto más importante, más digital.
- Estas correlaciones son moderadas, por lo que hay casos excepcionalmente positivos (como Zamacois o Valdés) y excepcionalmente negativos (como Antonio Machado o Miguel Hernández). Puede haber otros factores afectando la digitalización, como el principal género literario del autor.
- Ambas tendencias pueden guiar el trabajo de investigadores y grupos de trabajo tanto de manera positiva (intentar ser oportunista y trabajar en aquellos autores como Galdós o Baroja que están suficientemente digitalizados) como de manera negativa (encontrar autores importantes cuya obra no haya sido digitalizada, como los mencionados en el anterior punto). Las Humanidades Digitales se encontrarán con dificultades serias en su investigación sobre literatura cuyos autores murieron en los años 40 o posteriormente.

Confío en que este artículo permita tener una imagen más definida de la digitalización de la Edad de Plata y que algunas de sus conclusiones e ideas puedan ser reutilizadas por la comunidad interesada en metodologías de *distant reading*. La sencilla metodología aquí explicada y utilizada para analizar el estado de la digitalización puede ser reutilizada para tener una imagen de otras épocas de la literatura española que están recibiendo notable

atención por las Humanidades Digitales como la Edad Media o el Siglo de Oro. El trabajo podría ampliarse hasta disponer de datos sobre la literatura en español a lo largo del tiempo y en los diferentes países hispanohablantes. Además, estos datos pueden completarse históricamente para datar el estado de la digitalización de la literatura durante los siguientes años y décadas. Estos podrían ser algunos de los pasos para suplir la falta de textos literarios accesibles en formatos estándares internacionales, situación que esperamos solventar como comunidad de investigación.

REFERENCIAS BIBLIOGRÁFICAS

AGENJO, X. (2015). "Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos". *Ínsula: Revista de Letras y Ciencias Humanas*, 822, 12-15.

ALLÉS TORRENT, S. (2015). "Edición digital y algunas tecnologías aliadas". *Ínsula: Revista de Letras y Ciencias Humanas*, 822, 18-21.

BLEI, D.M. (2012). "Probabilistic Topic Models". *Communication of the ACM*, *55*.4, 77-84.

BURROWS, J. (2002). "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing*, 17.3, 267-287.

CALVO TELLO, J. (2016). "Entendiendo Delta desde las Humanidades". *Caracteres. Estudios culturales y críticos de la esfera digital*, 5.1, 140-176.

CALVO TELLO, J. y CEREZO SOLER, J. (en prensa). "La conquista de Jerusalén ¿de Cervantes? Análisis estilométrico sobre autoría en el teatro del Siglo de Oro español". *Digital Humanities Quarterly*, 10.

DE LA ROSA, J. y SUÁREZ, J.L. (2016). "The Life of *Lazarillo de Tormes* and of his Machine Learning Adversities Non-Traditional Authorship Attribution Techniques in the Context of the *Lazarillo*". *Lemir*, 20, 373-438.

EDER, M., KESTEMONT, M. y RYBICKI, J. (2016). "Stylometry with R: A Package for Computational Text Analysis". *The R Journal*, 1*6*.1, 1-15.

ERTLER, K.D. (2013). *Moralischen Wochenschriften*. Graz: Universität Graz. Recuperado de http://gams.uni-graz.at/archive/– el 29/04/2017.

EVANS, J.D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Brooks Cole Pub. Co.

GÓMEZ, S., CALVO TELLO, J., GONZÁLEZ, J.M. y VILCHES, R. (2015). "Hacia una biblioteca electrónica textual del teatro en español de 1868-1936 (BETTE)". *Texto digital*, 11.2, 171-184.

HASLWANTER, T. (2016). *An Introduction to Statistics with Python*. Cham: Springer International Publishing.

JAURALDE POU, P. (2012). *Clásicos Hispánicos*. Würzburg: More than Books. Recuperado de http://clasicoshispanicos.com/ el 28/04/2017.

JOCKERS, M.L. (2013). *Macroanalysis-Digital Methods and Literary History*. Champaign: University of Illinois Press.

LUCÍA MEGÍAS, J.M. (2012). *Elogio del texto digital. Claves para interpretar el cambio de paradigma*. Madrid: Fórcola Ediciones.

MAINER, J.C. (1981). *La edad de plata (1902-1939). Ensayo de interpretación de un proceso cultural*. Madrid: Cátedra.

MORETTI, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

NAVARRO-COLORADO, B. (2015a). "A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects". En *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, 105-113. Recuperado de http://www.aclweb.org/anthology/W15-0712 el 28/04/2017.

____ (2015b). *Corpus of Spanish Golden-Age Sonnets*. Valencia: Universidad de Alicante. Recuperado de http://github.com/bncolorado/CorpusSonetosSigloDeOro el 28/04/2017.

OLEZA SIMÓ, J. (dir.) (2013). *Biblioteca Digital Artelope*. Valencia: Universitat de València. Recuperado de http://artelope.uv.es/biblioteca el 28/04/2017.

RIßLER-PIPKA, N. (2016). "Avellaneda y los problemas de la identificación del autor. Propuestas para una investigación con nuevas herramientas digitales". En *El otro Quijote. La continuación de Avellaneda y sus efectos,* H. Ehrlicher (ed.), 27-51. Augsburg: Universität Augsburg.

SÁNCHEZ-MARTÍNEZ, F. (2013). *IMPACT-es Diachronic Corpus*. Alicante: Universidad de Alicante. Recuperado de https://www.digitisation.eu/tools-resources/language-resources/impact-es/ el 28/04/2017.

SCHÖCH, Ch. (2013). "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of the Digital Humanities*, 2.3, 2-13.

____ (2015). *Textbox*. Würzburg: Universidad de Würzburg. Recuperado de http://github.com/cligs/textbox el 28/04/2017.

SCHÖCH, Ch., HENNY, U. y CALVO TELLO, J. (2014). *The CLiGS toolbox*. Würzburg: University of Würzburg. Recuperado de https://github.com/cligs/toolbox el 28/04/2017.

SEMLAK, M. (2014). "Digitale Edition als Instrument für literaturwissenschaftliche Forschung". En *Literaturwissenschaft im digitalen Medienwandel*. 36-48.

SEVERIN, D. (2007). *An Electronic Corpus of 15th Century Castilian Cancionero Manuscripts*. Liverpool. Recuperado de http://cancionerovirtual.liv.ac.uk/ el 28/04/2017.

SIMÓN PALMER, M.C. (1997). *Teatro Español del Siglo de Oro*. Ann Arbor: ProQuest. Recuperado de http://teso.chadwyck.com/ el 27/04/2017.

TRILCKE, P., FISCHER, F. y KAMPKASPAR, D. (2015). "Digitale Netzwerkanalyse dramatischer Texte". En *DHd-Tagung*. Graz. Recuperado de http://gams.uni-graz.at/o:dhd2015.v.040 el 29/04/2017.

URRUTIA CÁRDENAS, S.H. (1999). "La Edad de Plata de la literatura española (1868-1936)". *Cauce: Revista de filología y su didáctica*, 22-23, 581-595.

ANEXO

Se puede acceder a la hoja de cálculo sobre el estado de la digitalización de la Edad de Plata desde http://www.caicyt-conicet.gov.ar/micrositios/hd/?attachment_id=1056