

**UNIVERSIDAD DE  
MURCIA**



FACULTAD DE COMUNICACIÓN Y DOCUMENTACIÓN

**GRADO EN INFORMACIÓN Y DOCUMENTACIÓN**

*TRABAJO FIN DE GRADO*

**ANÁLISIS DE WIKISOURCE COMO PLATAFORMA COMPLEMENTARIA PARA  
LA DIFUSIÓN DE LAS COLECCIONES DIGITALES DE BIBLIOTECAS**

AUTOR:

LAURA MORENO ZAMBRANO

REALIZADO BAJO LA TUTELA DEL PROFESOR:

TOMÁS SAORÍN PÉREZ

CONVOCATORIA JULIO 2017

# Índice

RESUMEN.....	3
ABSTRACT.....	4
1. INTRODUCCIÓN.....	5
1.1 OBJETIVOS DEL TRABAJO.....	5
2. DIGITALIZACIÓN DE PATRIMONIO BIBLIOGRÁFICO Y DOCUMENTAL.....	6
3. EL RECONOCIMIENTO ÓPTICO DE CARACTERES Y LA TRANSCRIPCIÓN DE DOCUMENTOS.....	10
4. TRANSCRIPCIÓN Y CORRECCIÓN COLABORATIVA.....	13
4.1 CROWDSOURCING Y MULTITUD.....	13
4.2 CROWDSOURCING EN LAS INSTITUCIONES CULTURALES.....	16
4.2.1 Transcripción colaborativa.....	16
4.2.2 Análisis de proyectos de transcripción .....	17
4.2.2.1 Transcriu-me!! .....	18
4.2.2.2 Transcrire .....	18
4.2.2.3 Europeana Transcribe .....	19
4.2.2.4 Trove .....	21
4.2.2.5 Operation War Diary .....	22
5. LOS PROYECTOS DEL ECOSISTEMA WIKIMEDIA.....	22
5.1. GLAM-WIKI.....	23
5.2. WIKISOURCE, EL PROYECTO DE TEXTOS FUENTE.....	24
5.3. WIKIMEDIA COMMONS, BANCO DE IMÁGENES Y MEDIOS.....	25
5.4. WIKIDATA, BANCO DE DATOS CENTRALIZADO.....	26
5.5. WIKIMEDIA Y LAS BIBLIOTECAS DIGITALES.....	26
6. ANÁLISIS PRÁCTICO DEL FUNCIONAMIENTO DE WIKISOURCE PARA LA TRANSCRIPCIÓN DE OBRAS DIGITALIZADAS.....	28
6.1 Criterios de selección.....	28
6.2. Trabajo en Wikisource.....	32
6.2.1 Depósito del fichero digitalizado: Wikimedia Commons.....	33
6.2.2 Integración de las obras digitalizadas: Wikisource.....	37
6.2.2.1 Formateo, transcripción y corrección.....	39
6.2.2.2 Transclusión o publicación de textos.....	42
6.2.2.3 Catalogación.....	44
7. CONCLUSIONES.....	48
REFERENCIAS BIBLIOGRÁFICAS.....	50
ANEXOS.....	56

## **RESUMEN**

El trabajo se propone evaluar la utilidad, costes y formatos de participar en el proyecto Wikisource desde el área de bibliotecas digitales y patrimonio cultural digital de las instituciones culturales. El proyecto Wikisource forma parte de la familia de proyectos de la Fundación Wikimedia para el conocimiento libre, junto a los más conocidos como Wikipedia o Commons. Se trata de un archivo de fuentes primarias en todos los idiomas, con licencia libre o libre de derechos.

La existencia de proyectos colectivos y colaborativos abre posibilidades a las bibliotecas para el desarrollo de sus políticas de digitalización de documentos y su difusión. Frente a la línea de actuación predominante, orientada a los repositorios digitales propios conforme a los requisitos de Europeana, existen otras vías con diferentes enfoques técnicos y de producción. Wikisource puede ofrecer por un lado características técnicas no disponibles en los entornos propios de las bibliotecas, y por otro puede permitir establecer otro tipo de relación con los usuarios y grupos de interés, produciendo un impacto positivo en la visibilidad del patrimonio digital y en la consecución de los objetivos de servicio al usuario y difusión de las colecciones.

El trabajo se plantea servir como guía a los bibliotecarios para entender las peculiaridades del proyecto Wikisource, explorar la naturaleza de los trabajos técnicos de formateo y versionado necesarias para publicar en él, analizar los costes de producción y gestión, y facilitar el desarrollo de estrategias de mejora de la línea de actuación en biblioteca digital.

### **Palabras clave:**

Digitalización / Bibliotecas digitales / Transcripción / Crowdsourcing en bibliotecas / Wikisource / Wikimedia / Digitalización participativa / OCR

## **ABSTRACT**

This investigation aims to evaluate the usefulness, costs and formats of participating in the Wikisource project from the area of digital libraries and digital cultural heritage of cultural institutions. The Wikisource project is part of the family project of the Wikimedia Foundation for Free Knowledge, along with the well-known Wikipedia or Commons. Wikisource is a free content library of source texts available in all languages, with public domain or free licenses.

The existence of collective and collaborative projects opens up possibilities for libraries to develop their policies for digitizing documents and disseminating them. As opposed to the predominant line of action, oriented to owned digital repositories according to the requirements of Europeana, there are other routes with different technical and production approaches. Wikisource can offer, on the one hand, technical characteristics that are not available in the libraries' own environments, and on the other hand, it can allow to establish another type of relationship with users and interest groups, having a positive impact on the visibility of digital heritage and the achievement of objectives of service to the user and dissemination of collections.

The work is intended to guide librarians in understanding the peculiarities of the Wikisource project, to explore the nature of the technical formatting works and the version needed to publish it, analyze the costs of production and management, and facilitate the development of strategies of improvement of the line of action in digital libraries.

### **Key Words:**

Digitization / Digital libraries / Transcription / Crowdsourcing in libraries / Wikisource / Wikimedia / Participative digitization / OCR

## 1. INTRODUCCIÓN

El presente trabajo se propone evaluar la utilidad, costes y formatos de participar en el proyecto Wikisource desde el área de bibliotecas digitales y patrimonio cultural digital de las instituciones culturales. El proyecto Wikisource forma parte de la familia de proyectos de la Fundación Wikimedia para el conocimiento libre, junto a los más conocidos como Wikipedia o Commons. Se trata de un archivo de fuentes primarias en todos los idiomas, con licencia libre o libre de derechos.

Para la consecución de este objetivo se desarrolla un análisis práctico del funcionamiento de Wikisource para la transcripción de obras digitalizadas en el cual se va a trabajar a partir de una selección de documentos digitalizados para su posterior incorporación al proyecto. Este trabajo pretende servir de guía para entender las peculiaridades del proyecto Wikisource, explorando la naturaleza de los trabajos técnicos de formateo y versionado necesarias para publicar en él y analizando los costes de producción y gestión.

La existencia de proyectos colectivos y colaborativos abre nuevas posibilidades a las bibliotecas para el desarrollo de sus políticas de digitalización de documentos y su difusión. Frente a la línea de actuación predominante, orientada a los repositorios digitales propios conforme a los requisitos de Europea, existen otras vías con diferentes enfoques técnicos y de producción.

Wikisource puede ofrecer por un lado características técnicas no disponibles en los entornos propios de las bibliotecas, y por otro lado puede permitir otro tipo de relación con los usuarios y grupos de interés.

### 1.1 OBJETIVOS DEL TRABAJO

Para el desarrollo de este trabajo se plantea un objetivo general y unos objetivos específicos. El **objetivo general** de este trabajo es evaluar la utilidad, costes y formatos de participación en el proyecto Wikisource desde el área de bibliotecas digitales y patrimonio cultural digital de las instituciones culturales.

Se plantea asimismo los siguientes **objetivos específicos**:

- Servir como guía a los bibliotecarios para entender las peculiaridades del proyecto Wikisource.
- Explorar la naturaleza de los trabajos técnicos de formateo y versionado necesarias para

publicar en él.

- Analizar los costes de producción y gestión.
- Facilitar el desarrollo de estrategias de mejora de la línea de actuación de bibliotecas digitales.

Se formulan además las siguientes **hipótesis** que se deben afirmar o refutar:

1. Wikisource produce un impacto positivo en la visibilidad y difusión de las colecciones digitales complementario a las acciones llevadas a cabo desde las instituciones culturales, que se traduce en un coste y unos beneficios para éstas.
2. Wikisource permite lograr otro tipo de relación con los usuarios y grupos de interés que resultan en la consecución de objetivos de servicio al usuario.

## **2. DIGITALIZACIÓN DE PATRIMONIO BIBLIOGRÁFICO Y DOCUMENTAL**

Las bibliotecas y otras instituciones culturales han desarrollado, a lo largo de su historia, métodos y herramientas para la transmisión de documentos a *texto completo* a través de diferentes medios. En el contexto actual, con las tecnologías de reproducción digital disponibles, las bibliotecas vuelven a dedicar medios y atención al uso, conservación y difusión del patrimonio documental y bibliográfico. Al hacerlo, se embarcan en el rol de producción y configuración de patrimonio cultural, sumándose al rol de preservación y difusión atribuido históricamente a éstas.

Según Dahlström y otros (2012) la digitalización se sumaría a la tradición que iniciaron actividades de transmisión de la cultura como la transcripción de manuscritos o, más recientemente, la microfilmación. Lo que tiene en común con los métodos que le han precedido en la historia es su finalidad, que puede resumirse en acceso y preservación (Carrera, 2012).

Los proyectos de digitalización han formado parte de la actividad de las instituciones culturales desde hace décadas. Con inicio en los años ochenta, las primeras experiencias en grandes bibliotecas datan de la década de los noventa. Actualmente, los proyectos se han extendido y la digitalización se ha convertido progresivamente en una práctica establecida en bibliotecas, archivos y otras instituciones culturales. Las instituciones se encuentran inmersas en un proceso

generalizado de digitalización de patrimonio cultural, especialmente de patrimonio bibliográfico, así como de las colecciones de audio, vídeo, fotografías, cuadros, mapas, etc. Pero junto a la digitalización sigue presente la problemática del acceso y difusión en el medio digital (Robledano, 2016):

El patrimonio ya digitalizado es inmenso, pero a la hora de acceder podemos hablar de dispersión: tenemos miles de sistemas de repositorio que conforman un inmenso conglomerado de "islas de información" con no mucha conexión entre sí. La dispersión supone una seria barrera para el acceso al patrimonio ya digitalizado para muchos ciudadanos y para su control de cara a establecer políticas de digitalización. Una de las vías de trabajo es unificar el acceso y control mediante la creación de bibliotecas virtuales interinstitucionales y transnacionales.

Los primeros esfuerzos de digitalización de patrimonio bibliográfico se centraron únicamente en una cuidada selección de materiales o colecciones singulares. Se trataba de proyectos que, en la mayoría de los casos, alcanzaban una proyección limitada y se veían condicionados por las necesidades concretas de los materiales seleccionados (Bremner, 2011).

Si bien este planteamiento resultaba válido para actividades como la transcripción de manuscritos o la microfilmación, surge el peligro de que los materiales digitalizados se encuentren aislados; fuera de contextos nacionales e internacionales.

Para asegurar la sostenibilidad de los materiales digitales se hace necesario el establecimiento de políticas y estándares internacionales. A nivel europeo, Europeana es el resultado a gran escala de este enfoque. En el proyecto Europeana se integran los objetivos económicos y culturales de toda Europa para la creación de un patrimonio cultural europeo común (Arquero y Cobo, 2014).

En el ámbito de la digitalización, las acciones europeas que han culminado en el proyecto Europeana han supuesto un punto de inflexión, al alcanzarse tanto masa crítica, conciencia pública, tecnologías eficientes, cooperación institucional y resultados tangibles. El informe de 2011 denominado "Nuevo Renacimiento" situaba la digitalización como estrategia básica en la política digital cultural de la Unión Europea (Comisión Europea, 2011).

Europeana<sup>1</sup> es el portal que provee el acceso al patrimonio cultural y científico de Europa.

---

1 Disponible en <<http://www.europeana.eu/portal/es>>.

Lanzado por la Comisión Europea en el año 2008, ofrece actualmente acceso a libros, material de audio y vídeo, fotografías, cuadros, mapas, manuscritos, periódicos y documentos de archivo digitalizados que forman parte del patrimonio cultural de Europa.

Europeana funciona como un portal multimedia con contenidos provenientes de diferentes fuentes. Los objetos digitales accesibles a través del portal no se almacenan en éste, sino que permanecen en las instituciones culturales que custodian los contenidos. Europeana recopila información contextual acerca de los objetos y proporciona acceso al contenido completo disponible en los servidores de las respectivas instituciones.

Los datos ofrecidos por el informe de *Estrategia Europea 2015-2020* señalan que el 90% del patrimonio europeo está aún por digitalizar. Ese 10% digitalizado representa alrededor de 300 millones de objetos que reflejan las diversas facetas de la cultura europea: libros, cuadros, cartas, fotografías, sonido e imagen en movimiento. Sin embargo, sólo un tercio (34%) de estos objetos está disponible en línea, y únicamente un 3% es reutilizable (*Europeana Strategy 2015-2020*, s.f.).

Una vez puesta a disposición una masa crítica de contenido, la publicación con el único objetivo de aumentar la cantidad de contenido ya no es una prioridad para Europeana y las instituciones que trabajan con ella. La prioridad número uno actualmente es la mejora de metadatos y la calidad del contenido (*Europeana Foundation, 2017*). Se apuesta por una estrategia de adquisición y publicación de contenido basada en la demanda de los usuarios.

La actualización de la *Estrategia Europea 2020* introduce nuevas recomendaciones para originar un cambio en los contenidos que tenga en cuenta a los usuarios, sus necesidades y sus demandas. Entre ellas se encuentra el desarrollo de colecciones temáticas atractivas, así como de temáticas que animen a la interacción a través de los diferentes países europeos. Se anima además a valorar el feedback de los usuarios para desarrollar colecciones de calidad que estén basadas en demandas.

Por último, llama a las instituciones a alentar la participación de los usuarios para enriquecer y mejorar los datos a través de eventos relacionados con el usuario y la literatura, como los *Transcribathons* o maratones de transcripción. Se anima a que bibliotecas, archivos y museos trabajen con el usuario para aumentar la calidad de los datos y ampliar el compromiso público

con las instituciones de una forma innovadora.

Si bien los proyectos de digitalización se encontraron con una primera fase de incertidumbre, donde predominaba la individualidad, el aislamiento y la falta de interconexión, las experiencias auguran un futuro hacia el uso de estándares que implican el acceso global y la preservación del patrimonio. La puesta a disposición, divulgación y explotación del patrimonio cultural va más allá de marcos nacionales e institucionales. En este contexto, surgen problemáticas comunes que todavía hoy imperan en el ámbito de la digitalización y que se deben solventar a nivel nacional e internacional.

Arquero y Cobo (2014) destacan tres desafíos a los que deben hacer frente las instituciones culturales:

1. Los derechos de autor,
2. La disponibilidad de infraestructuras tecnológicas que sustenten el proyecto,
3. Y el compromiso de los gobiernos a asumir parte de los costes derivados de estos proyectos.

El régimen actual de derechos de autor es muy problemático para la digitalización masiva y no se adapta a las necesidades de acceso transfronterizas, fundamental en el ecosistema en que coexisten estas colecciones digitales.

Algunas instituciones están creando nuevos derechos al digitalizar obras de dominio público<sup>2</sup> (Comisión Europea, 2011). Desde el informe del *Comité de Sabios, grupo de reflexión para poner en línea el patrimonio cultural de Europa*, sin embargo, se apunta a que "el material de dominio público digitalizado con fondos públicos debe ser accesible gratuitamente a todos". Es decir, no se debe generar nuevos derechos sobre materiales que ocupan la categoría de dominio público, y se debe unificar las prácticas de las instituciones en cuanto a derechos de uso. Asimismo, este material debe estar disponible de manera gratuita para su reutilización no comercial. Esto implica además (Llueca, 2016) dejar de emplear marcas de agua y otros elementos que puedan limitar el acceso a los documentos y los metadatos que se encuentren

---

2 COMMUNIA, 2010. "El dominio público es la riqueza de información que está libre de barreras de acceso o de reutilización usualmente asociada a la protección de la propiedad intelectual, ya sea porque está libre de cualquier protección de derechos o porque los titulares de derechos han decidido eliminar dichas barreras". <<http://www.publicdomainmanifesto.org/manifesto>>

en dominio público. En cuanto a las condiciones de acceso a este material, se ha establecido un consenso para que, una vez haya sido digitalizado, éste se haga accesible sin restricciones.

Las instituciones culturales comenzaron digitalizando sus colecciones sin derechos de autor por razones económicas y prácticas (Comisión Europea, 2011). La negociación de contratos y la liberación de derechos de autor supondría, en la gran mayoría de los casos, gastos imposibles de asumir por parte de las instituciones. El patrimonio cultural más reciente está sujeto a derechos de autor, y las bibliotecas deben digitalizarlo para asegurar su preservación. Pero su difusión sin restricciones es una tarea más compleja (Pereira, 2014). Se destaca la ausencia de obras accesibles del siglo XX, el *agujero negro* producido por el hecho de que buena parte de las obras clásicas no son digitales, están fuera de distribución y a menudo son huérfanas<sup>3</sup> (Serra, 2011). Al respecto, el Comité recomienda crear un instrumento legal europeo que las regule, evitar que se sigan generando obras huérfanas en el futuro mediante la creación de algún tipo de registro previo como condición para ejercer los derechos de autor, y promover el uso de licencias colectivas.

### **3. EL RECONOCIMIENTO ÓPTICO DE CARACTERES Y LA TRANSCRIPCIÓN DE DOCUMENTOS**

El objetivo de integrar plenamente el patrimonio cultural en el entorno de las tecnologías de la información y la comunicación sólo puede lograrse mediante la digitalización a texto completo (Dahlström y otros, 2012). La digitalización de patrimonio bibliográfico no se reduce a la conversión de imágenes a formato digital, sino que supone la obtención de verdaderos corpus de texto que permiten la recuperación y difusión de la información para la generación de nuevo conocimiento.

Algunas de las ventajas de disponer de las transcripciones de documentos en formato digital son la preservación de la información textual, la disponibilidad de los datos y la capacidad para realizar búsquedas y traducciones de forma automática (Bocanegra, 2016; Cordal, 2016). Permite además dar impulso a nuevas investigaciones, como los estudios históricos o léxicos, partiendo del contenido disponible en hemerotecas y bibliotecas digitales.

---

3 Real Decreto 224/2016, de 27 de mayo, por el que se desarrolla el régimen jurídico de las obras huérfanas, BOE núm. 141 § 5717 (2016). "Obra cuyos titulares de derechos de propiedad intelectual no están identificados o, de estarlo, no están localizados a pesar de haberse efectuado una previa búsqueda diligente de los mismos".

El reconocimiento óptico de caracteres u OCR (Optical Character Recognition) es el sistema que permite reconocer el texto escaneado o fotografiado en una imagen digital y generar un archivo en formato de texto legible por ordenador. Es decir, a través de estas herramientas se logra un transcripción automática de texto: partiendo una imagen digital que contiene texto, se obtiene su representación textual digital.

El sistema OCR funciona realizando un análisis de una imagen digital y dividiendo esa imagen en componentes estructurales más pequeños para encontrar zonas de contenido textual. Dentro de estas zonas se identifican líneas individuales de texto y, dentro de esas líneas, se identifican caracteres y palabras individuales. Una vez identificado un carácter, compara sus propiedades con las fuentes de texto disponibles para encontrar una coincidencia. Repite el proceso para todos los caracteres de una palabra y los compara con un diccionario de palabras completas para encontrar una coincidencia. Este proceso se desarrolla con oraciones, líneas y bloques de texto hasta que se identifica todo el texto de una imagen digital.

La transcripción es la acción de escribir un texto en otro formato bien como simple representación de un documento, o bien como datos de investigación que pueden ser incorporados a bases de datos (Alcalá, 2015). Se trata de un proceso esencialmente humano que requiere a veces conocimiento experto: ha sido durante mucho tiempo objeto preferente de interés para amplios colectivos científicos como historiadores, paleógrafos y diplomatas, filólogos o documentalistas (López, 1998).

La transcripción de documentos resulta un terreno propicio para estudiar desde las *humanidades digitales*, campo de estudio sobre el que se distinguen varios enfoques: desde las tecnologías aplicadas a disciplinas encuadradas en las humanidades tradicionales hasta nuevos campos de investigación como los estudios de internet (Rodríguez-Yunta, 2013).

A nivel internacional, IFLA e ICA (2014) en su texto *Directrices para proyectos de digitalización de colecciones y fondos de dominio público, en particular para aquellos custodiados en bibliotecas y archivos* señala “facilitar nuevas formas de acceso y uso” de los colecciones como una de las principales razones para la digitalización. Destaca la posibilidad de acceso que abre para todo tipo de materiales “de un modo más fácil y productivo utilizando herramientas

avanzadas de ordenador como OCR (Reconocimiento Óptico de Caracteres) o codificación del texto para textos convertidos”.

Desde estas Directrices se recomienda “ofrecer imágenes en color en alta resolución [...] y, de ser posible, la opción de buscar en el texto, transcripciones completas o parciales del texto”.

Dispone a la libre elección de las instituciones si se utiliza un procedimiento de digitalización que reproduzca la imagen o se utiliza OCR o se transcribe el texto original. IFLA e ICA (2014):

Es probable que los usuarios quieran textos recuperables, y esto significa OCR o teclado [...]. Por otro lado, dependiendo del tipo de usuarios y del tipo de texto muchos usuarios querrán también ver las imágenes de la página, y la experiencia del aspecto del original. Esto puede llevarnos a la conclusión de usar ambos métodos aunque en la mayoría de los casos eso puede suponer un coste prohibitivo. Así, la mejor solución es elegir las imágenes de la página.

A nivel nacional, la Subdirección General de Coordinación Bibliotecaria, en *Requisitos técnicos de los proyectos de digitalización de patrimonio bibliográfico y de prensa histórica de la SGCB*, establece como exigencia el procesado de OCR para las publicaciones seriadas. En cuanto a patrimonio bibliográfico, sólo recomienda el procesado y el uso de programas de gestión de repositorios o bibliotecas digitales que permitan la búsqueda en el texto en los proyectos subvencionados por esta Subdirección. Ministerio de Educación, Cultura y Deporte (2016):

En los contratos de digitalización de la Subdirección General de Coordinación Bibliotecaria se exigirá procesado de OCR (Optical Character Recognition) para las páginas digitalizadas de publicaciones seriadas. Se exige la entrega de un archivo OCR en formato ALTO (Analyzed Layout and Text Object) por cada página digitalizada. Estos archivos OCR hacen posible la búsqueda en el texto de las páginas digitalizadas en la Biblioteca Virtual de Prensa Histórica.

El proceso de OCR depende de una serie de factores que influyen de manera significativa en la calidad de los resultados. El factor más relevante es la naturaleza de los documentos que se han de digitalizar, sumada a su posterior conversión a imagen digital. Ballesteros, Morales y Cedill (2012) indican que para lograr resultados positivos en el proceso de OCR es necesario que, en primer lugar, el texto en el documento original sea “legible, exento de roturas o manchas, con letras uniformes y bien impresas”. Por otra parte, la representación digital que se obtenga de éste debe ser “nítida, encuadrada, sin perspectiva o deformaciones por curvatura y a una

resolución suficiente que permita la captura fiel del texto impreso”. Otros factores que pueden afectar al resultado final del OCR (National Library of Australia, s.f.) son el tratamiento previo que se dé a la imagen, los formatos de imágenes sobre los que se trabaje, los diccionarios de los que disponga el software para el proceso de comparación de caracteres o la capacidad de estos programas para *recordar* el modo de actuar frente a problemáticas reiteradas.

El procesado de los documentos que reúnen estas características dan excelentes resultados y presentan índices muy altos de éxito en la búsqueda y recuperación de texto. Sin embargo, si alguno de estos factores carece de precisión, afectará negativamente al proceso del OCR, y la combinación de estos factores disminuirá aún más la precisión. Para materiales bibliográficos más antiguos y prensa histórica, el rendimiento del OCR presenta problemas: a menudo es inconsistente y los textos resultantes no son fieles al contenido textual impreso.

Disponer de textos recuperables supone un coste significativo dentro de un proyecto de digitalización y las instituciones deberán decidir si el uso del OCR es la forma más conveniente de poner sus colecciones a disposición. En algunos proyectos de digitalización se opta por la transcripción manual de texto, pero este proceso resulta más costoso ya que requiere de una gran mano de obra. En ocasiones se valorará la combinación de OCR y transcripción manual para poder disponer de textos completos.

#### **4. TRANSCRIPCIÓN Y CORRECCIÓN COLABORATIVA**

La producción cultural es, en los últimos tiempos (Roig, Sánchez y Leibovitz, 2017), un sector muy proclive a iniciativas basadas en la implicación de una multitud interesada en proyectos o en procesos de creación colectiva.

Para entender el *boom* de la creación colectiva hay que señalar el ejemplo esencial que representa Wikipedia. Aunque no es el primer caso en la práctica de la colaboración masiva, se considera el paradigma del fenómeno *crowdsourcing* (Serra, 2017).

##### **4.1 CROWDSOURCING Y MULTITUD**

Para entender el fenómeno de la transcripción colaborativa, primero hay que conocer conceptos como inteligencia colectiva y crowdsourcing.

Segaran (2007) define la *inteligencia colectiva* como la combinación de comportamientos, preferencias o ideas de un grupo de personas que genera una nueva perspectiva. Si bien el concepto y los métodos utilizados para aprovechar la inteligencia colectiva son anteriores a la existencia de internet, la capacidad de recopilar información de miles o millones de personas a través de internet ha abierto muchas nuevas posibilidades. Por su parte, Lévy (citado en Cobo, 2007) señaló, ya en los años noventa, la necesidad de crear "herramientas tecnológicas que permitieran la construcción cooperativa de conocimientos de muchos con muchos". El autor defendía que, si las tecnologías se orientaban a ser mediadoras entre las inteligencias de los individuos de la sociedad, éstos podrían ver potenciadas sus capacidades creativas.

El término *crowdsourcing*, popularizado a partir de un artículo de Jeff Howe de la revista Wired en 2006 (Estellés y González, 2012a; Roig et al., 2017), se refiere a un modelo de resolución de tareas que implica la participación de una multitud en internet. Las iniciativas crowdsourcing representan una de las muchas maneras de aprovechar la inteligencia colectiva (Estellés, Navarro y González, 2015).

Howe (citado en Roig et al., 2017) definía en un primer momento crowdsourcing como "el proceso en el que una empresa o institución traslada una tarea que tradicionalmente era llevada a cabo por sus trabajadores a un conjunto de personas externas de número indefinido pero habitualmente grande a través de una convocatoria abierta".

A lo largo de los años esta definición ha ido cambiando y adaptándose a nuevas realidades. No es necesariamente una empresa o institución quien inicia el proceso, y la externalización de tareas no nace exclusivamente para aliviar las cargas de trabajo de empresas e instituciones.

Por ejemplo, Roig y otros (2017) hablan de un "proceso de colaboración", donde bien un individuo o bien una organización inicia el proceso. Ya se habla de "resolución de un problema o superación de un reto" y no de externalización de tareas. Igualmente, Estellés y González (2012b) hablan de "iniciativas de tipo participativo". Estos autores destacan además su relación con fenómenos como la innovación abierta o la inteligencia colectiva.

Debido a la amplitud del concepto y la falta de base teórica se desarrollan estudios que, a partir de diferentes enfoques metodológicos, presentan nuevas visiones sobre el concepto de crowdsourcing. Estellés y González (2012a) proponen una clasificación de iniciativas de crowdsourcing basada en "la tarea a realizar" por la multitud como criterio diferenciador.

Basándose en este criterio, los autores distinguen entre los siguientes tipos:

- Crowdcasting.
- Crowdcollaboration.
- Crowdcontent.
- Crowdfunding.
- Crowdopinion.

El crowdcasting es un llamamiento para la resolución de un problema donde se ofrecen recompensas a quien lo resuelva antes o de la manera más efectiva. En el crowdcollaboration la multitud aporta su conocimiento para resolver problemas o plantear ideas, pero no se ofrecen recompensas. El crowdfunding busca financiación mediante pequeñas aportaciones y en el crowdopinion se intenta conocer la opinión de la multitud sobre un tema o un producto.

Por su parte, las iniciativas de **crowdcontent** destacan por la aportación de mano de obra y conocimiento de una multitud que crea o encuentra contenido de diversa naturaleza. Cada individuo trabaja de manera individual y luego se reúne el resultado de todos (Estellés y González, 2012a). Para los autores, la *crowdproduction* o *producción de la multitud* queda englobada a su vez bajo esta denominación, y hace referencia a la creación de contenido de manera colaborativa a través de tareas de dificultad variable. Se trata de la iniciativa con una relación más directa y que se encuentra más vinculada al mundo de bibliotecas, archivos y museos (Alcalá, 2015), por lo que cabría incluir la transcripciones colaborativas masivas dentro de las iniciativas crowdcontent.

Por otra parte, el estudio de Roig y otros (2017) plantea una noción de crowdsourcing más amplia, de forma que su definición acoja, entre otros:

- "Proyectos nativos que tienen en el crowdsourcing su razón de ser.
- Iniciativas que se orientan a la continuidad y la replicación a lo largo del tiempo.
- Valoración del impacto y resultados de la actividad fruto del crowdsourcing: construcción de relevancia y desarrollo de confianza."

Andro (2016) va más allá y plantea hablar de *communitysourcing* o *nichesourcing* en lugar de crowdsourcing para los proyectos participativos de bibliotecas digitales, haciendo referencia al

sentido de *comunidad* frente a una multitud anónima. Aunque estas iniciativas se dirigen teóricamente a un número ilimitado de usuarios, se ha comprobado que, para la mayoría de los proyectos, la mayor parte de los datos producidos son aportados por una pequeña minoría de los participantes definida (el 80% del trabajo se haría por sólo el 10% de los voluntarios más activos) y no una multitud.

## **4.2 CROWDSOURCING EN LAS INSTITUCIONES CULTURALES**

Desde hace unos años, las bibliotecas y otras instituciones culturales están llevando a cabo iniciativas crowdsourcing. Algunas de estas iniciativas están relacionadas (Colmenero-Ruíz, 2015) con la contextualización de objetos, la complementación de colecciones, la clasificación de objetos, la co-gestión de datos y la corrección y transcripción de documentos. Dobrecky (2016) señala otras iniciativas como la identificación de imágenes, el etiquetado de documentos, la detección de errores en los catálogos, la agregación de información a los registros, la creación de libros electrónicos o la descripción de material.

Queda establecido el campo de estudio sobre iniciativas crowdsourcing impulsadas desde las bibliotecas. Bibliografía como la tesis doctoral *Bibliothèques numériques et crowdsourcing* (Andro, 2016) sobre transcripción y corrección en bibliotecas digitales, o las recientes ponencias *Colaboración abierta en las instituciones de la memoria* (Serra, 2017) y *Wikipedia en las instituciones de la memoria: nuevos territorios, nuevas costumbres* (Saorín, 2017) demuestran que se está asentando una base teórica en torno a los procesos y proyectos de colaboración abierta que se están desarrollando en las instituciones culturales.

### **4.2.1 Transcripción colaborativa**

La escritura es un fenómeno que tiene más de cinco mil años. Ésta supone el fin de la prehistoria y el nacimiento de las grandes civilizaciones del mundo antiguo, las cuales conocemos hoy gracias a sus textos (García, 2016). La evolución de la escritura y el valor que se otorga a los documentos textuales están relacionados con la forma en que se difunden los tipos de escritura: en la actualidad, los dispositivos electrónicos han cambiado nuestra forma de comunicarnos, de leer, de escribir y de almacenar mensajes u otros tipos de contenidos (Martínez, 2016).

En este contexto, la transcripción de textos presenta una nueva dimensión: necesitamos más que nunca poder relacionarnos en el entorno digital con las fuentes textuales que bibliotecas, archivos y otras instituciones están poniendo a disposición a través de los esfuerzos de la digitalización.

Muchos de los proyectos de transcripción colaborativa nacen con el objetivo de generar "información descriptiva y textual que pueda servir de punto de acceso rastreable en documentos históricos", y sirven a las instituciones culturales para resolver limitaciones (de costes y personal) en la comprobación y corrección de grandes cantidades de texto, mejorar el acceso a las colecciones y establecer relaciones directas con los usuarios (Neudecker y Tzadok, 2010; Alcalá, 2015).

Brumfield (2015) anima a que las instituciones inmersas en procesos de digitalización de patrimonio documental y bibliográfico expongan estos esfuerzos de digitalización en proceso al usuario mediante iniciativas de transcripción colaborativa. Este autor defiende que, al exponer a los usuarios a los flujos de trabajo, los materiales pueden conectar de forma directa con los intereses de los usuarios.

Desde las instituciones culturales se debe apostar por iniciativas de transcripción colaborativa que no se encuentren limitadas a una tarea o un período concretos. La noción de que una multitud colabore en la transcripción y corrección de documentos puede ser un fin en sí mismo. Debe asegurarse además la perdurabilidad y relevancia de estas iniciativas en el tiempo. Es importante que la institución se plantee unos objetivos y se evalúen los resultados de forma periódica, y que se mantenga relevante a los intereses y las expectativas del público al que sirve.

#### **4.2.2 Análisis de proyectos de transcripción**

Se analiza a continuación cinco proyectos de transcripción impulsados por instituciones culturales relevantes a nivel nacional e internacional.

Para ello se ha recogido datos sobre:

- La colección: criterios de selección, volumen de ésta.
- Resultados de transcripción y evaluación del proceso.
- El uso de herramientas y la plataforma sobre la que se trabaja.

#### 4.2.2.1 Transcriu-me!!

[\(http://transcriu.bnc.cat/\)](http://transcriu.bnc.cat/)

La colección de este proyecto catalán se basa en fondos de la Biblioteca Catalana y, además, también forman parte de él los fondos de la Filmoteca de Cataluña y la Universidad de Barcelona. Actualmente, las colecciones que se encuentran disponibles en la página y susceptibles de transcripción son cuatro (todos están finalizados): la obra literaria de Àngel Guimerà, Testigos bibliotecarios, Juntas Literarias del Real Colegio de Cirugía de Barcelona y los Pergaminos de Organyà.

**Transcriu-me!!** És una iniciativa de col·laboració (crowdsourcing) impulsada per la Biblioteca de Catalunya per millorar l'accés als continguts digitals. És oberta a la participació d'altres institucions. Les biblioteques i institucions digitalitzen documents històrics manuscrits o difícils d'interpretar automàticament, i vosaltres els transcriu en línia.

Les transcripcions acabades s'incorporen periòdicament a les digitalitzacions de la *Memòria Digital de Catalunya*; d'aquesta manera, gràcies a vosaltres, els documents es poden cercar per qualsevol paraula del text, fet que millora els resultats de les cerques al portal, facilita el descobriment dels documents i contribueix a fer conèixer el patrimoni del país.

**Testimonis bibliotecaris** (Col·lecció amb transcripcions pendents)

**Obra literària d'Àngel Guimerà** (Col·lecció amb totes les transcripcions finalitzades)

Generalitat de Catalunya  
Departament de Cultura

Biblioteca de Catalunya

La página es de fácil acceso al usuario ya que no tiene que realizar ningún tipo de registro o ingreso a la plataforma. De igual manera, informa sobre el estado de las transcripciones y ofrece la posibilidad de ver el histórico de transcripciones de años anteriores, aportando además el período temporal que la obra necesitó para ser totalmente transcrita y un gráfico de evaluación. Todos estos documentos digitalizados transcritos migran tras este proceso a la Memoria Digital de Cataluña.

El proyecto se desarrolla gracias a la plataforma Omeka, sobre la que se integra un plugin script para realizar las transcripciones y software MediaWiki para la gestión de usuarios, permisos y transcripciones.

#### 4.2.2.2 Transcrire

[\(http://transcrire.huma-num.fr/\)](http://transcrire.huma-num.fr/)

El proyecto TRANSCRIRE centra la selección de fondos en colecciones de cuadernos, libros, folletos o colecciones de antropólogos, arqueólogos, historiadores y geógrafos franceses. El proyecto obtiene sus archivos de las colecciones de varios socios de las bibliotecas de investigación en Francia como son la biblioteca Yves Le Gallo de Breton, el Centro de Investigación celta (CRBCh), el archivo de sonido de la casa mediterránea de Ciencias Humanas (Sr. MSH) o la biblioteca Geografía de Sorbonne (Paris 1, 4 y 7), entre otros.

Actualmente disponen seis colecciones para transcribir: los cuadernos de campo Raymond Mauny, Cuadernos de campo de Marceau Gast, Cuadernos Annie-Hélène Dufour, Periódicos y colección de Anatole Le Braz, Libro de Emmanuel de Martonne y los Folletos de François Chevalier. Al entrar en las páginas individuales de cada uno nos ofrece una breve descripción del cuaderno o el folleto, para, a continuación, presentar las miniaturas de los documentos adjuntando el porcentaje de transcripción que tiene cada uno.



The screenshot shows the TRANSCRIRE website interface. At the top, there is a dark navigation bar with the logo 'TRANSCRIRE' in green and white, and links for 'TRANSCRIRE PAR COLLECTION', 'SE CONNECTER', and 'A PROPOS'. Below the navigation bar, the page title is 'Cahiers de terrain de Raymond Mauny'. A short biography of Raymond Mauny (1912-1994) is provided, stating he was a French historian and specialist in prehistory and protohistory of Africa, with notebooks from 1942-1962. Below the text, there are five columns, each representing a different collection of notebooks. Each column includes a thumbnail image of a notebook cover, a title and date, and a progress bar indicating the percentage of transcription completed. The collections are: 'Mission Gao Soudan (Mali), 29 janvier 1950 - 10 février 1950' (2% débuté), 'Mission au Soudan (Mali), novembre 1954 - février 1955' (4% débuté), 'Mission El Beyyed (Adrar de Mauritanie), Noël 1958' (5% débuté), 'Cahier de Préhistoire Cap Vert (Sénégal), 1946' (6% débuté), and 'Mission Sénégal Soudan Mauritanie, 8 au 30 janvier 1948' (9% débuté).

Collection	Percentage
Mission Gao Soudan (Mali), 29 janvier 1950 - 10 février 1950	2% débuté
Mission au Soudan (Mali), novembre 1954 - février 1955	4% débuté
Mission El Beyyed (Adrar de Mauritanie), Noël 1958	5% débuté
Cahier de Préhistoire Cap Vert (Sénégal), 1946	6% débuté
Mission Sénégal Soudan Mauritanie, 8 au 30 janvier 1948	9% débuté

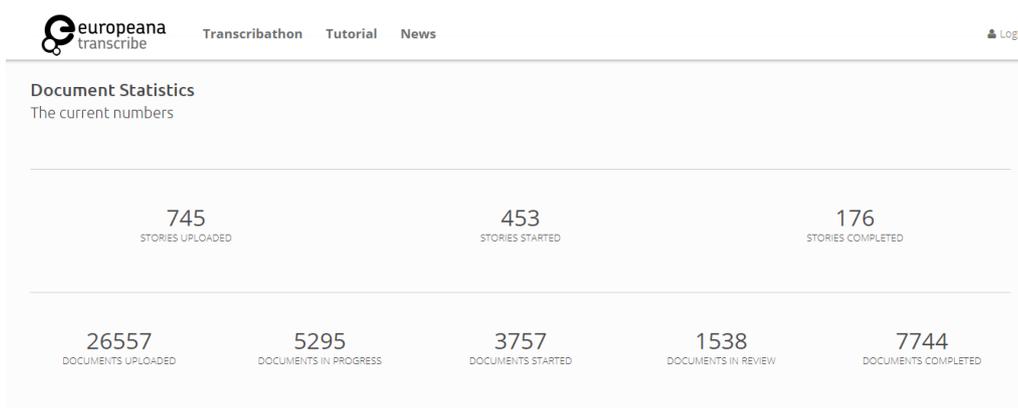
La página web del proyecto tiene un diseño sencillo e intuitivo. A diferencia de la anterior en ésta sí debemos estar registrados para comenzar a transcribir los documentos. A nivel de plataforma, este proyecto, al igual que el catalán, hace uso de la plataforma Omeka con el mismo plugin script y el software MediaWiki.

#### 4.2.2.3 Europeana Transcribe

(<https://transcribathon.com/en/>)

El proyecto Europeana Transcribe centra sus colecciones en documentos de multitud de puntos del mundo alrededor de la Primera Guerra Mundial, fuentes que se encuentran asociadas a la página Europeana 1914-1918. Obtiene así recursos de colecciones nacionales, archivos cinematográficos y registros inéditos.

El volumen de documentos con los que cuenta el proyecto es de 26.557 documentos, de los cuales 5.294 se encuentran en proceso de transcripción y 7.744 están ya finalizados. La página cuenta con una pestaña donde se pueden comprobar los resultados de manera pormenorizada, incluyendo el número de usuarios y sus categorías.



Cuenta además con un mapa de puntos calientes donde, escogiendo por los diferentes países, aparecen los documentos que aportan al proyecto y el estado de transcripción en el que se encuentran. Esta misma información se ofrece para el estado de los documentos en proceso de transcripción. Al posar el cursor sobre las miniaturas de los archivos aparece una ficha que ofrece información como el número de páginas del documento, el porcentaje completado de esas páginas y los usuarios que están trabajando en él.



La plataforma es accesible e intuitiva. En ella también debemos registrarnos para comenzar a transcribir. Una de las cosas que destacamos de la página es la posibilidad que ofrece Europea de participar en grupos de transcripción, así como la realización de maratones o campañas que añaden un aspecto lúdico a los procesos buying de transcripción.

#### 4.2.2.4 Trove

(<http://trove.nla.gov.au/>)

El proyecto TROVE es impulsado por la colaboración de la Biblioteca Nacional Australiana, las bibliotecas estatales y territoriales, así como cientos de instituciones culturales y de investigación de Australia. Cuenta con más de 536.457.331 de recursos entre libros, imágenes, periódicos históricos, mapas, música y diferentes archivos.

La página actúa en primera instancia como un gran repositorio donde vienen a parar todos los documentos de las instituciones australianas anteriormente citadas. No podemos encontrar un icono que nos lleve directamente a documentos para transcribir debido a que, para cualquier archivo susceptible de ser transcrito, se abrirá automáticamente el menú para ello.



Para el volumen de documentos que trabaja el portal la búsqueda de documentos se hace fácil y sencilla al usuario. En cuanto a la transcripción la página ofrece la posibilidad de acceder a la transcripción ingresando un código captcha sin necesidad de registrarse.

El proyecto ofrece un histórico del volumen de documentos añadidos, la cantidad de diarios o periódicos que han sido digitalizados y el número de usuarios que participan en la iniciativa.

#### 4.2.2.5 Operation War Diary

(<https://www.operationwardiary.org/>)

Por último, el proyecto Operation War Diary reúne documentos originales de la Primera Guerra Mundial del Archivo Nacional Británico, así como la visión histórica del Museo Imperial de la Guerra y los datos de la comunidad de Zooniverse<sup>4</sup>.

Este proyecto tiene alrededor de 1,5 millones de páginas de diarios de guerra, los cuales cubren la actividad en el frente occidental. La página aunque intuitiva en un primer momento, obliga al usuario a registrarse y a realizar un tutorial paso por paso.

Sobre los diarios, el documento fundamental de trabajo, ofrece información a nivel histórico y sobre los datos del nivel de transcripción.



## 5. LOS PROYECTOS DEL ECOSISTEMA WIKIMEDIA

Más allá de Wikipedia existen más sitios web basados en el modelo wiki para la edición de contenidos, pero el éxito de Wikipedia puede hacer que se confunda el caso con el modelo. Sin embargo, se trata tan sólo de la aplicación de la tecnología wiki en un proyecto de propósito muy específico a gran escala, junto a una serie de reglas de funcionamiento de una comunidad online. Wikipedia supone diminuta parte de internet que surge como resultado de una idea innovadora (la colaboración *radical*) y que se sostiene por el esfuerzo continuado de miles de voluntarios en todo el planeta (Saorín, 2013). El éxito y la presencia cotidiana de Wikipedia nos puede hacer perder la perspectiva para comprender el funcionamiento de un proyecto de

---

4 Zooniverse es un portal de ciencia ciudadana que permite a los usuarios participar en la investigación científica a través de la aportación de datos reales.

carácter singular, usado como ejemplo de éxito para apoyar la viabilidad de usar la “inteligencia colectiva” o “crowdsourcing” en el medio digital.

Wikipedia tiene varias iniciativas hermanas que complementan la enciclopedia. Todas ellas son multilingües, libres, apoyadas en la tecnología wiki y con infraestructura de la Fundación Wikimedia (Wikimedia Foundation). Podemos hablar de la existencia de un *ecosistema Wikimedia*. Cada proyecto hermano tiene sus peculiaridades (Commons, Wikisource, Wikicitas). En estos proyectos wiki conviene enfocar nuestra atención hacia dos ámbitos diferenciados (Saorín, 2012): los contenidos y las personas; la comunidad que edita y las reglas a las que se suscriben.

La Fundación Wikimedia ha prestado mucha atención a los estudios generales de usuarios, tratando de extraer conclusiones útiles de la demografía y comportamiento de los editores, con el fin de afrontar la ampliación y mantenimiento de la masa de participación.

Conocer la realidad multilingüe de los proyectos Wikimedia es necesario para comprender éstos. Los proyectos son muchos y éstos son diferentes e independientes (Saorín, 2012):

La amplitud y riqueza de entradas de cada versión es sintomática de la actividad digital de la comunidad correspondiente. En tanto esto es así, podría ser un indicador más de desarrollo de la sociedad de la información o de participación digital de cada comunidad lingüística.

Algunas limitaciones y problemáticas de un proyecto colaborativo como Wikipedia puede ser:

- No alcanzar suficiente masa crítica de participación y contenidos para adquirir relevancia.
- El desequilibrio en la aportación de los participantes: pocos usuarios activos.
- Dificultad de recompensar por el *esfuerzo* de colaboración.

### **5.1. GLAM-WIKI**

Aunque Wikipedia surge desde la iniciativa social, desde las organizaciones sociales públicas y privadas también se puede aportar valor de forma directa e indirecta. Cada vez más instituciones entienden que pueden cumplir mejor su función social dedicando una parte de sus recursos a la construcción del procomún digital, del que Wikipedia es uno de sus ejemplos más señalados (Saorín, 2013).

Las instituciones podrán desarrollar modelos de participación coherentes con sus recursos y actividades. Desde *Wikimedia Outreach* se recopila y evalúa buenas prácticas para ampliar el alcance del proyecto Wikimedia mediante la colaboración con el ecosistema cultural real (Saorín, 2012). Una de sus áreas clave se denomina **GLAM** (Galleries, Libraries, Archives and Museums)<sup>5</sup>. La conferencia GLAM-Wiki de 2009 elaboró una serie de recomendaciones que tratan de ajustar las concepciones de las instituciones culturales con la cultura abierta, en aspectos como la financiación de proyectos, los derechos morales y de reproducción o las tecnologías.

Un documento que confirma la importancia de esta colaboración es el presentado en el congreso de IFLA de 2016 sobre *Oportunidades para bibliotecas académicas y de investigación y Wikipedia* (IFLA, 2016).

## 5.2. WIKISOURCE, EL PROYECTO DE TEXTOS FUENTE

Wikisource se centra en construir un archivo wiki de textos libres de fuentes primarias en cualquier idioma, así como de traducciones de textos. Originalmente se llamó Proyecto Sourceberg, pero en 2003 la comunidad adoptó la denominación actual Wikisource. Su evolución es en cierto modo la inversa al banco de imágenes y medios Commons, del que trataremos después: empezó como un depósito centralizado y durante su crecimiento, en 2005, se crearon versiones específicas por idiomas. Describimos aquí de forma genérica los principios y funcionamiento de la edición en español de Wikisource: <http://es.wikisource.org>

En Wikisource los editores se dedican a proporcionar los textos fuente de forma precisa y accesible. Contiene tanto libros publicados y documentos históricos, así como trabajos de fuentes notables y otros trabajos con referencias. Responde, por lo tanto, a una jerarquía de prioridades a la hora de seleccionar los textos fuente a subir al repositorio, análoga a la política de "notabilidad enciclopédica" de Wikipedia.

La comunidad ha establecido una serie de principios para el trabajo en Wikisource, que abarcan aspectos de contenido y comunidad:

- Precisión. Una serie de criterios para mantener la fidelidad al texto original digitalizado y transcrito.

---

5 "El agrupar estas actividades obedece a la observación de algunos analistas del hecho de que estas instituciones tienen en común como forma de operar la creación de colecciones, y que muchas veces se solapan". Wikimedia España/Proyectos/GLAM. <[https://www.wikilovesearth.es/eo/Wikimedia\\_Espa%C3%B1a/Proyectos/GLAM](https://www.wikilovesearth.es/eo/Wikimedia_Espa%C3%B1a/Proyectos/GLAM)>.

- Accesibilidad. Lectura en cualquier plataforma, por lo que la maquetación se realiza aplicando opciones para un diseño multidispositivo.
- Colaboración. El objetivo esencial de la colaboración ha de ser la calidad de los textos, y no otros aspectos secundarios de formato o presentación.
- Formato. En la medida de la posible hay que buscar mantener la fidelidad a la disposición y organización del texto original, aunque sea de forma aproximada.
- Convivencia. Este principio tiene una denominación poco clara, pero hace referencia a no incluir demasiadas mejoras ajenas al texto original al transformarlo en una serie de páginas wiki.

La herramienta de gestión del sitio web permite la carga de ficheros digitalizados, en conexión con Commons, la realización de OCR, la gestión colaborativa del proceso de transcripción y validación, y el seguimiento de las diferentes etapas de revisión (Texto en progreso; texto completado; revisado y corregido; revisado por varios usuarios) y ciclo de vida. Una de las peculiaridades de Wikisource es que los textos, una vez revisados y corregidos por la comunidad, se protegen, evitando nuevas ediciones.

### **5.3. WIKIMEDIA COMMONS, BANCO DE IMÁGENES Y MEDIOS**

Wikimedia Commons es un proyecto de la Fundación Wikimedia creado en 2004 con el objetivo de servir como repositorio de archivos de imágenes y otros formatos multimedia para el resto de los proyectos. Como sucede con otros proyectos de la Fundación, todos sus archivos deben encontrarse en dominio público o haber sido liberados bajo licencias libres. Se basa en un wiki configurado de tal forma que los archivos allí contenidos están disponibles en todos los demás wikis de los proyectos Wikimedia: esto permite una gestión centralizada de estos archivos en todas las versiones de Wikipedia, Wikisource, etc.

El banco de imágenes Wikimedia Commons forma, junto con los artículos enciclopédicos, la pareja con más éxito de los proyectos de Wikimedia, en cuanto a uso, coherencia y crecimiento (Pastor-Sánchez y Saorín, 2011). Los contenidos de Commons son reutilizables en cualquier contexto, siempre que se mantenga el mismo tipo de licencia abierta con la que fueron depositados. En este contexto, las bibliotecas son las encargadas de aportar materiales únicos de enorme trascendencia: no basta la digitalización de documentos, debe preverse y

asegurarse la compatibilidad de licencias Creative Commons o Dominio Público para que puedan ser usados sin limitaciones.

#### **5.4. WIKIDATA, BANCO DE DATOS CENTRALIZADO**

Wikidata es la base de datos secundaria libre, colaborativa y multilingüe de la Fundación Wikimedia. Wikidata recopila datos estructurados para dar soporte a Wikisource, Wikipedia, Wikimedia Commons y otros proyectos Wikimedia.

Este proyecto se basa en varios pilares fundamentales:

- **Colaboración.** Al igual que los demás proyectos Wikimedia, los datos de Wikidata son agregados y mantenidos por editores del proyecto, quienes deciden las normas de creación y de gestión de contenidos.
- **Multilingüismo.** La edición, la explotación, la navegación y la reutilización de los datos es totalmente multilingüe. Los datos introducidos en cualquier idioma están disponibles de forma inmediata en todos los demás idiomas.
- **Estructuración de datos.** Wikidata recoge datos en forma estructurada, lo que permite a los proyectos Wikimedia y a terceras partes reutilizar fácilmente esos datos, facilitando además el procesado por parte de ordenadores.
- **Verificabilidad.** Se trata de una base de datos secundaria: Wikidata no sólo guarda la información registrada, sino también las fuentes de las que ésta procede.

Estamos ante un paso más para entender otra dimensión de la organización colaborativa de conocimiento: la formalización colaborativa.

#### **5.5. WIKIMEDIA Y LAS BIBLIOTECAS DIGITALES**

Aunque el personal de las bibliotecas pueda participar en los proyectos Wikimedia a título individual, la participación como institución debe formar parte de una línea de acción consciente, evaluable y alineada con el resto de actuaciones en el marco de las estrategias digitales globales (Saorín, 2012).

El informe *Report on the results of the Wikimedia Taskforce* (Europeana, 2015) establece una serie de recomendaciones aplicables a bibliotecas digitales, entre las que destacan:

- Se debe proporcionar ayuda para facilitar las conexiones locales entre GLAM y *Wikimedians*.
- Convertir a Wikidata en un elemento central de la estrategia de la plataforma Europea.
- Europea debe seguir invirtiendo en tecnología que mejore la interoperabilidad entre GLAMs y plataformas de Wikimedia.

Europeana concede cada vez más importancia en sus estrategias al impacto en el medio digital, y no sólo a la cantidad de objetos digitalizados en bibliotecas digitales. La mera puesta a disposición de un gran catálogo no implica que sea altamente visible para el ciudadano (Saorín y Rodríguez, 2012). Ante el temor de las instituciones culturales por la pérdida de visibilidad de sus colecciones en el espacio digital, estos autores plantean la elaboración de nuevas métricas para medir el impacto de sus acciones en el medio digital, que en lugar de medir simplemente el tráfico en sus sitios web, tengan además en cuenta "la generación de accesos en sitios ya frecuentados por los usuarios".

Los responsables de proyectos digitales pueden potenciar su misión y resultados a través de la incorporación de sus colecciones en los proyectos Wikimedia. En el caso de Wikisource, la Biblioteca Nacional de Francia, la Biblioteca Central de Florencia o el Archivo Nacional de Estados Unidos presentan ejemplos de colaboración desde bibliotecas y archivos:

- Wikitrento.  
[https://outreach.wikimedia.org/wiki/GLAM/Newsletter/November\\_2015/Contents/Italy\\_report](https://outreach.wikimedia.org/wiki/GLAM/Newsletter/November_2015/Contents/Italy_report)
- Archives départementales des Alpes-Maritimes.  
<http://manuscripttranscription.blogspot.com.es/search/label/mediawiki>
- US National Archives and Records Administration.  
[https://outreach.wikimedia.org/wiki/GLAM/Model\\_projects/Improving\\_the\\_quality\\_of\\_OCR](https://outreach.wikimedia.org/wiki/GLAM/Model_projects/Improving_the_quality_of_OCR)
- Biblioteca Nazionale Centrale di Firenze.  
[https://it.wikipedia.org/wiki/Progetto:GLAM/Biblioteca\\_Nazionale\\_Centrale\\_di\\_Firenze](https://it.wikipedia.org/wiki/Progetto:GLAM/Biblioteca_Nazionale_Centrale_di_Firenze)

## **6. ANÁLISIS PRÁCTICO DEL FUNCIONAMIENTO DE WIKISOURCE PARA LA TRANSCRIPCIÓN DE OBRAS DIGITALIZADAS**

Partimos en el inicio de la investigación de la hipótesis de que Wikisource produce un impacto positivo en la visibilidad y difusión de las colecciones digitales complementario a las acciones llevadas a cabo desde las instituciones culturales, que se traduce en un coste y unos beneficios para éstas. Y de que permite, además, lograr otro tipo de relación con los usuarios y grupos de interés que resultan en la consecución de objetivos de servicio al usuario.

Para demostrar o refutar esta hipótesis, al inicio de este trabajo se planteaban los siguientes objetivos específicos a los que se pretende dar respuesta con la elaboración de este trabajo:

- Servir como guía a los bibliotecarios para entender las peculiaridades del proyecto Wikisource.
- Explorar la naturaleza de los trabajos técnicos de formateo y versionado necesarias para publicar en él.
- Analizar los costes de producción y gestión.
- Facilitar el desarrollo de estrategias de mejora en la línea de actuación de bibliotecas digitales.

Para la realización del trabajo de campo se ha partido de una primera fase de selección de documentos digitalizados, para incorporarlos posteriormente al proyecto Wikisource en una segunda fase, de forma que se pueda evaluar la utilidad, los costes y los formatos de participación en este proyecto. A continuación se detalla ambas etapas del proceso.

### **6.1 Criterios de selección**

Desde este trabajo de campo se parte de la selección de obras ya digitalizadas que se encuentran disponibles en bibliotecas digitales. A continuación se señalan los criterios que se han tenido en cuenta para la selección de obras.

En primer lugar, como requisito esencial, las obras seleccionadas debían cumplir las condiciones que permitieran su libre utilización. Para su utilización (edición, reproducción o difusión pública) se deben cumplir dos condiciones (Real Decreto Legislativo 1/1996):

1. Por una parte, las obras deben estar en dominio público. Una obra pertenece al dominio

público cuando se extinguen los derechos de explotación sobre ésta.<sup>6</sup> Estos derechos de explotación duran toda la vida del autor y setenta años después de su muerte o declaración de fallecimiento.<sup>7</sup> No obstante, el plazo es de ochenta años para los autores fallecidos antes del 7 de diciembre de 1987 (Biblioteca Nacional de España, 2017).

2. En segundo lugar, debe respetarse la autoría y la integridad de la obra. Para esta segunda condición, se aseguró que para todas las obras utilizadas durante la realización del trabajo se señalara de forma clara y concisa su autoría.

Por lo tanto, el primer criterio que se tuvo en cuenta fue que las obras se encontraran en dominio público al haber vencido de forma efectiva los derechos de explotación. Esto es, para el año 2017, aquellos autores fallecidos hasta el año 1936. A modo de ejemplo, en el año 2017 se incorporan al dominio público (Biblioteca Nacional de España, 2017b) las obras de autores como Federico García Lorca, Ramón María del Valle-Inclán o Miguel de Unamuno.

Posteriormente, se acotó la búsqueda a obras de autores murcianos. Lorente (2016) destaca la importancia de las colecciones locales como proveedoras de materiales inéditos y valiosos para conocer la historia y las costumbres de una localidad pero afirma que, a menudo, muchas necesidades específicas de información de temas de ámbito local no se pueden satisfacer en internet debido a la escasez de estos materiales en el medio digital. Saorín (2011), por su parte, defiende que, aunque los contenidos de interés local surgen y son consumidos en un entorno limitado, "la naturaleza relacional de la red permite que se integren con fuentes de información de escala global".

El motivo de la elección de autores murcianos cuyos materiales se encontraban en los distintas bibliotecas digitales de la Región es su desvinculación de estos repositorios y del ámbito regional para integrarlos a un proyecto a escala mundial.

En este paso resultó especialmente útil la búsqueda en distintas bibliotecas digitales de la Región de Murcia. Destacan por sus colecciones la Biblioteca Digital de la Región de Murcia, la Biblioteca Digital Floridablanca y la Biblioteca Digital de la Ciudad de Murcia.

---

6 Real Decreto Legislativo 1/1996. TÍTULO IV. Dominio público. Artículo 41. Condiciones para la utilización de las obras en dominio público.

7 Real Decreto Legislativo 1/1996. CAPÍTULO I. Duración. Artículo 26. Duración y cómputo.

- La Biblioteca Digital de la Región de Murcia<sup>8</sup>, impulsada por la Consejería de Cultura y Turismo y gestionada desde la Biblioteca Regional, compila toda clase de documentos relacionados con la Región, ya sea por su temática, impresión o autoría murcianas.
- La Biblioteca Digital Floridablanca<sup>9</sup>, por su parte, tiene como objetivo permitir el acceso abierto del Fondo Histórico de la Universidad de Murcia, facilitando su consulta, su difusión y su preservación.
- Por último, la Biblioteca Digital de la Ciudad de Murcia<sup>10</sup>, conserva, preserva y difunde el fondo bibliográfico que se conserva en el Archivo Municipal de Murcia.

### 6.1.1 Selección de obras

A la hora de seleccionar los documentos se consideraron dos criterios diferenciadores: en primer lugar, la distinción entre documentos impresos y manuscritos. Y, en segundo lugar, entre documentos que incorporaran reconocimiento óptico de caracteres y documentos que no lo hicieran.

De esta forma, se pretende comprobar si los resultados varían según el tipo de documento y la presencia o no de OCR. Se pretende además valorar la variación que estas características producen en los procesos de transcripción colaborativa en el contexto del proyecto Wikisource.

Por lo tanto, se trabaja sobre los siguientes documentos:

1. Manuscrito con reconocimiento óptico de caracteres  
Pedro Jara Carillo. *Chirigotas* (1918).  
Biblioteca Digital de la Región de Murcia.
2. Manuscrito sin reconocimiento óptico de caracteres  
Ricardo Sánchez Madrigal. *A su reja* (18--).  
Biblioteca Digital de Murcia.
3. Impreso con reconocimiento óptico de caracteres  
Pedro Jara Carillo. *Paco Cayuela* (1901).  
Biblioteca Digital de la Región de Murcia.
4. Impreso sin reconocimiento óptico de caracteres  
Ricardo Codorníu. *Doce árboles* (1914).  
Biblioteca Digital Floridablanca.

8 Disponible en <<http://bibliotecadigital.carm.es/inicio/index.php>>.

9 Disponible en <<https://bibliotecafloridablanca.um.es/bibliotecafloridablanca/>>.

10 Disponible en <<http://www.murcia.es/jspui/>>.

**Ricardo Sánchez Madrigal** (1844-1925) fue ingeniero de minas, escritor, periodista y poeta. Se dedicó a la ingeniería, actividad que compaginó con la dirección del periódico *La Verdad* y con su labor creativa como poeta y autor de obras de teatro.

Obra: *A su reja*

Manuscrito sin reconocimiento óptico de caracteres

Poema

Fecha de publicación: 18--

Descripción física: [2] h. ; 4º

Recuperado de: Biblioteca Digital de Murcia<sup>11</sup>

**Pedro Jara Carrillo** (1876-1927) fue poeta, novelista, autor de cuentos y de obras de teatro, periodista y político.

Obra: *Chirigotas*

Manuscrito con reconocimiento óptico de caracteres

Poema

Fecha de publicación: 1918

Descripción física: 2 h.

Recuperado de: Biblioteca Digital de la Región de Murcia<sup>12</sup>

Obra: *Paco Cayuela*

Impreso con reconocimiento óptico de caracteres

Monólogo teatral

Fecha de publicación: 1901

Descripción física: XV, 22 p., [2] h.

Recuperado de: Biblioteca Digital de la Región de Murcia<sup>13</sup>

**Ricardo Codorniu y Stárico** (1846-1923) fue ingeniero de montes y autor de una amplia obra científica y técnica centrada en la difusión y enseñanza del medio forestal.

Obra: *Doce árboles*

Impreso sin reconocimiento óptico de caracteres

Libro de relatos

Fecha de publicación: 1914

Descripción física: 64 p.

---

11 Ejemplar disponible en <<http://hdl.handle.net/10645/1169>>.

12 Ejemplar disponible en <<http://bibliotecadigital.carm.es/opac/ficha.php?informatico=00006276>>.

13 Ejemplar disponible en <<http://bibliotecadigital.carm.es/opac/ficha.php?informatico=00001342>>.

Por último, resulta significativo mencionar que algunas instituciones al digitalizar están generando nuevos derechos sobre materiales que ocupan la categoría de dominio público, imposibilitando su utilización. Además, las marcas de agua y otros elementos limitan asimismo la reutilización de los documentos. Durante la realización de este trabajo se descartó la selección de las bibliotecas digitales Biblioteca Digital Hispánica<sup>15</sup> y Biblioteca Digital Gredos<sup>16</sup> al generar a través de la digitalización nuevos derechos sobre materiales de dominio público, así como hacer uso de marcas de agua.

## 6.2. Trabajo en Wikisource

Esta fase del trabajo de campo está dividida en dos pasos. Por una parte, el depósito del fichero digitalizado. Y, posteriormente, la integración de los documentos digitalizados en el proyecto.

Para la agregación de textos en Wikisource existen dos vías: de forma preferente, a través de un archivo digitalizado del texto (generalmente en formato PDF o DjVu) a partir del cual se realizará su transcripción. Además, existe la opción de transcribir directamente un texto en caso de no disponer de su versión digitalizada.

En cualquiera de estos dos casos, previamente los textos deben reunir una serie de requisitos para asegurar su incorporación y mantenimiento en Wikisource:

1. El documento debe encontrarse en Dominio Público o debe disponer de una Licencia Creative Commons Atribución-CompartirIgual 3.0 o compatible (2.0, 4.0, GFDL) que permita:
  1. Compartir — copiar y redistribuir el material en cualquier medio o formato.
  2. Adaptar — remezclar, transformar y crear a partir del material.
2. Se debe determinar la fecha de la publicación/edición de la obra.
3. Se debe determinar su autor.
4. Para las obras originarias no escritas en castellano, debe conocerse el traductor de la

---

14 Ejemplar disponible en <<http://hdl.handle.net/11169/670>>.

15 La Biblioteca Digital Hispánica es la biblioteca digital de la Biblioteca Nacional de España. Disponible en <<http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html>>.

16 La Biblioteca Digital Gredos forma parte del repositorio documental de la Universidad de Salamanca. Disponible en <<https://gredos.usal.es/jspui/handle/10366/1>>.

obra y la fecha de publicación de la edición traducida. Es importante señalar aquí que aunque una obra se encuentre en dominio público (o licencias similares) las traducciones de ésta pueden no estarlo y es fundamental disponer de esta información.

5. Debe señalarse la **fuentes** desde donde se ha conseguido el archivo digitalizado o se ha transcrito el texto. En este punto se asegura el reconocimiento de la autoría y la integridad de una obra, al mismo tiempo que se reconoce y señala la institución que conserva los materiales.

### 6.2.1 Depósito del fichero digitalizado: Wikimedia Commons

Aunque todo el proceso de trabajo se va a realizar sobre Wikisource, las imágenes no se encontrarán almacenadas en esta página, sino que deben integrarse en el repositorio multimedia Wikimedia Commons. El primer paso es el depósito del fichero digitalizado en Wikimedia Commons.

Si no queremos desplazarnos a Wikimedia Commons, desde la página principal de Wikisource se enlaza, en su sección de *herramientas*, con el asistente de subidas de Commons.

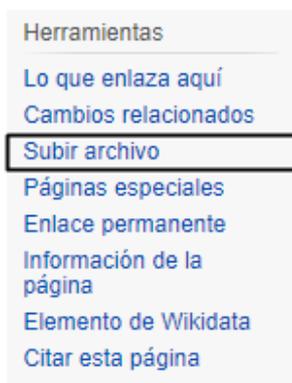


Figura 1. Opción de subir archivo en Wikisource.

En esta primera fase de inserción de los archivos digitalizados tienen lugar tres procesos:

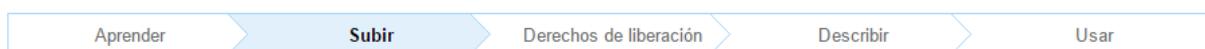


Figura 2. Asistente de subidas de Wikimedia Commons: proceso de subida guiado.

### 1. Subida de archivo en formato PDF o DjVu.

Las tres bibliotecas digitales seleccionadas disponen sus documentos en formato PDF.

### 2. Aportación de información sobre **derechos de liberación**, de forma que se asegure su reutilización legal.

Este repositorio acepta la incorporación de obras creadas por terceros con la condición de que los derechos de explotación de la obra permitan su utilización, copia, modificación y uso comercial. Para este tipo de documentos debe especificarse la fuente desde donde se obtiene el documento (la biblioteca digital que lo custodia), los autores (en este paso se ha optado por referenciar la institución encargada de la digitalización) y el régimen de los derechos de autor. En este último punto varían las condiciones para su incorporación a dominio público que establecía la legislación española (ochenta años desde la muerte de un autor), ya que el repositorio está alojado en Estados Unidos y la obra debe estar libre de derechos de autor en ese país.

Wikimedia Commons permite la subida de imágenes de obras publicadas por primera vez antes de 1923 y cuyo autor haya fallecido hace más de setenta años.

#### ● Este archivo no es mi propio trabajo.

Fuente \*



[<http://bibliotecadigital.carm.es/opac/ficha.php?informativo=00006276MO&idpag=1244416194&presenta=digitalty&codopac=OP042> Biblioteca Regional de Murcia. JARA 418]

Autor(es) \*



[<http://bibliotecadigital.carm.es/inicio/index.php> Biblioteca Digital de la Región de Murcia]

Ahora especifica por qué estás seguro de que tienes derecho a publicar este trabajo:

- ▶ [El propietario de los derechos de autor publicó esta obra con la correcta licencia de Creative Commons correcta](#)
- ▶ [El propietario de los derechos de autor publicó su foto o vídeo en Flickr con la licencia correcta](#)
- ▼ [Los derechos de autor han caducado definitivamente en los Estados Unidos](#)
  - Wikimedia Commons está alojado en los Estados Unidos, por lo que la obra debe estar libre de derechos de autor en ese país.
  - Publicado por primera vez en los Estados Unidos antes de 1923
  - Publicado por primera vez antes de 1923 y autor fallecido hace más de 70 años
  - Reproducción fidedigna de una pintura que está en el dominio público debido a su antigüedad

Figura 3. Información sobre derechos de explotación de la obra Chirigotas (1918) de Pedro Jara Carrillo.

### 3. Descripción del archivo.

Debe proporcionarse un título, una descripción detallada, una fecha de creación y una o

varias categorías. Estas categorías, características del software MediaWiki, permiten agrupar archivos multimedia relacionados a través del uso de etiquetas. Para categorizar los ficheros añadidos al repositorio se ha optado por hacerlo con los nombres de los autores de las obras. En el supuesto de que no exista una **categoría** con el nombre del autor de la obra en Commons, se debe crear una.

Título \*

Chirigotas

Descripción \*

español ▼ Poema manuscrito de Pedro Jara Carrillo publicado en 1918

[Agregar una descripción en otro idioma](#)

Fecha de creación \*

1918

Categorías

Pedro Jara Carrillo X Agregar una categoría...

Figura 4. Descripción de la obra Chirigotas (1918) de Pedro Jara Carrillo.

Así, para el autor Pedro Jara Carrillo se creó la categoría `[[Category:Pedro Jara Carrillo]]`, de forma que sus obras se agrupen bajo esta denominación. Al crear una nueva categoría se debe, a su vez, categorizar ésta. Para la categoría Pedro Jara Carrillo se ha optado por las siguientes:

`[[Category:People of the Region of Murcia]]`  
`[[Category:Writers from the Region of Murcia]]`

Una vez incluido el archivo en el repositorio, desde Wikimedia Commons se dispone de:

- Un resumen que identifica sus propiedades más significativas.

### Resumen [\[ editar \]](#)

<b>Descripción</b>	Español: Poema manuscrito de Pedro Jara Carrillo publicado en 1918
<b>Fecha</b>	1918
<b>Fuente</b>	Biblioteca Regional de Murcia. JARA 418 <a href="#">↗</a>
<b>Autor</b>	Biblioteca Digital de la Región de Murcia <a href="#">↗</a>

Figura 5. Visualización del resumen del archivo subido a Commons.

Para posteriores ediciones de estos datos, se debe trabajar directamente con el código que se generó automáticamente con el asistente de subidas.

Así, este resumen se materializa en:

```

=={{int:filedesc}}==
{{Information
|description={{es|1=Poema manuscrito de Pedro Jara Carrillo publicado en 1918}}
|date=1918
|source=[http://bibliotecadigital.carm.es/opac/ficha.php?
informatico=00006276MO&idpag=1244416194&presenta=digitalty&codopac=OP042 Biblioteca Regional
de Murcia. JARA 418]
|author=[http://bibliotecadigital.carm.es/inicio/index.php Biblioteca Digital de la Región de Murcia]
|permission=
|other versions=
}}

```

- La licencia a la que se acoge la obra.

#### Licencia [\[editar\]](#)



Este material está en **dominio público** en los demás países donde el derecho de autor se extiende por **70 años** (o menos) tras la muerte del autor.

Este archivo está en el **dominio público** en los Estados Unidos porque fue publicado antes del 1 de enero de 1923.

Esta obra ha sido identificada como libre de las restricciones conocidas en virtud del derecho de autor, incluyendo todos los derechos conexos.

Figura 6. Licencia de la obra Chirigotas.

Y su correspondiente código:

```

=={{int:license-header}}==
{{PD-old-70-1923}}

```

- El historial de modificaciones del archivo.
- Los usos del archivo en otros proyectos Wikimedia.
- Los metadatos del archivo, añadidos por la cámara digital o el escáner usado para su digitalización.

Software usado	Suite 102
Programa de conversión	EZTwainX 3,51b1 using EZPdf 1,60
Cifrado	no
Tamaño de página	<ul style="list-style-type: none"> <li>• 455,94 x 620,28 pts</li> <li>• 450,54 x 616,14 pts</li> </ul>
Versión del formato PDF	1,4
<a href="#">Ocultar datos detallados</a>	

Figura 7. Ejemplo de metadatos originados en el proceso de digitalización.

En Wikimedia Commons permanecen los objetos y éstos son susceptibles de incorporarse a cualquier proyecto de la Fundación Wikimedia. A modo de ejemplo, los distintos documentos que se han incorporado al repositorio durante la realización de este trabajo ya se encuentran enlazados con los artículos enciclopédicos de los autores en Wikipedia y se encuentran disponibles para ilustrar estos artículos.

### Enlaces externos [ editar código · editar ]

---

-  [Wikimedia Commons](#) alberga contenido multimedia sobre **Pedro Jara Carrillo**.
-  [Wikisource](#) contiene obras originales de o sobre **Pedro Jara Carrillo**.
- [Perfil de Pedro Jara Carrillo](#), por Manuel Llanos de los Reyes, en Tonos Digital, núm. 4, 2012.
- [Biografía de Jara Carrillo](#)
- [Biografía y obra de Pedro Jara Carrillo](#)
- [Perfil de Pedro Jara Carrillo](#)
- [Biografía de Pedro Jara Carrillo](#)

Figura 8. Sección de enlaces externos del artículo enciclopédico de Pedro Jara Carrillo.

Para enlazar otros proyectos con el artículo del autor en Wikipedia tan sólo es necesario incluir en él el siguiente código que remite, respectivamente, a su categoría en Commons y su ficha de autor (sobre las cuales hablaremos más adelante) en Wikisource:

```
{{commonscat|Pedro Jara Carrillo}}
```

```
{{wikisource}}
```

El objeto de Wikimedia Commons no es el acceso a una representación textual, sólo la puesta a disposición de la imagen digitalizada. El proyecto Wikisource trabaja con las imágenes disponibles en el repositorio y las incorpora a su flujo de trabajo.

#### 6.2.2 Integración de las obras digitalizadas: Wikisource

Tras la agregación de la imagen digitalizada a Wikimedia Commons, el siguiente paso es su integración en Wikisource.

El primer paso es la creación de una página dentro del espacio de nombres *Índice*, de modo que todos los documentos tengan un espacio dedicado en exclusiva al proceso de transcripción colaborativa. Los índices se dividen en cinco estados, de acuerdo al progreso de la transcripción de sus páginas o su naturaleza:

- Índices no corregidos.
- Índices corregidos.

- Índices validados.
- Índices dañados.
- Extractos.

Si el documento digitalizado que hemos subido a Wikimedia Commons recibía el nombre de "A su reja.pdf", a la hora de crear el índice de esta obra tan sólo es necesario darle a la nueva página el nombre "Índice:A su reja.pdf". Desde estos índices se muestran el archivo digitalizado, sus datos bibliográficos básicos y el enlace a cada una de las páginas que debe ser creada con su correspondiente transcripción textual.

Índice [Discusión](#) [Leer](#) [Editar](#) [Ver historial](#)

## Índice:A su reja.pdf



<b>Autor</b>	Ricardo Sánchez Madrigal
<b>Título</b>	A su reja: canción
<b>Año</b>	1874
<b>Imprenta o editor</b>	
<b>Fuente</b>	Biblioteca Digital de la Ciudad de Murcia - MU-AM, 10-E-15(15) <a href="#">@</a>
<b>Progreso</b>	Terminado — Todas las páginas validadas
<b>Páginas</b>	<a href="#">1</a> <a href="#">2</a> <a href="#">3</a> <a href="#">4</a> <a href="#">5</a> <a href="#">6</a> <a href="#">7</a> <a href="#">8</a> <a href="#">9</a> <a href="#">10</a> <a href="#">11</a> <a href="#">12</a> <a href="#">13</a>

Categoría: [Índices validados](#)

Figura 9. Índice de la obra A su reja.

Este Índice funciona a modo de carta de presentación del documento desde donde se navega hacia adelante y hacia atrás sobre las diferentes páginas que deben ser creadas.



Figura 10. Navegación sobre el Índice de un obra.

A la hora de crear estos Índices, Wikisource exige unos datos básicos que describan la obra e informen de su localización.

- Título: El título entero de la obra.

- Autor: El nombre del autor del libro, a ser posible con un enlace a su página.
- Traductor: El nombre del traductor de la obra, a ser posible con un enlace a su página.
- Imprenta o editor: Editorial, imprenta, etc.
- Ilustrador.
- Año de publicación.
- Fuente.

Estos datos bibliográficos se materializan en los siguientes campos:

<b>Título</b>	[[A su reja  A su reja: canción]]
<b>Autor</b>	[[Ricardo Sánchez Madrigal]]
<b>Año de publicación</b>	1874
<b>Fuente</b>	[ <a href="http://www.murcia.es/jspui/handle/10645/1169">http://www.murcia.es/jspui/handle/10645/1169</a> Biblioteca Digital de la Ciudad de Murcia - MU-AM, 10- E-15(15)]

Estos Índices presentan, además, elementos propios de Wikisource que permiten normalizar el proceso de transcripción:

<b>Imagen</b>	Número de la página con la imagen representativa de la obra. Por defecto "1".
<b>Progreso</b>	Progreso de transcripción de las páginas presentes en una obra. Valores: <i>Terminados</i> , <i>Por validar</i> , <i>Por corregir</i> , <i>Dañado</i> , <i>Extracto</i> .
<b>Páginas</b>	Páginas disponibles del archivo digitalizado. Por defecto "<pagelist />".
<b>Encabezado y pie de página</b>	En el caso de que haya encabezados y/o pies de página iguales en todas las páginas.

#### 6.2.2.1 Formateo, transcripción y corrección

Una vez integrados los ficheros digitalizados en Wikisource y tras la creación de los Índices de la obra, el siguiente paso es la creación de páginas. Veremos en este apartado cómo funcionan la transcripción y corrección colaborativa.

El primer paso es la creación de las distintas páginas que componen una obra. Como hemos visto en el punto anterior, al crear el Índice con la etiqueta "<pagelist />" se traspasan todas las imágenes que componían el documento PDF que se había depositado en Wikimedia Commons.

<b>Autor</b>	Pedro Jara Carrillo
<b>Título</b>	Chirigotas
<b>Año</b>	1918
<b>Imprenta o editor</b>	
<b>Fuente</b>	Biblioteca Regional de Murcia. JARA 418 <a href="#">↗</a>
<b>Progreso</b>	Por corregir
<b>Páginas</b>	1 2
	Página:Chirigotas.pdf/1 (la página no existe)

Figura 11. Estado de las páginas de la obra Chirigotas. El color rojo indica que no existen.

En las páginas de **transcripción** se trabaja con la versión digitalizada de los textos. Se trata de páginas que tienen en un lado la imagen de una página de un texto y en el otro su transcripción. Al momento de transcribir es importante dar formato al texto tal como aparece en la publicación, de manera que sea accesible. Es importante señalar que Wikisource usa el software MediaWiki y que trabajaremos para la elaboración de las transcripciones, entre otras páginas del proyecto, con wikitexto<sup>17</sup>.

(véase Anexo 1)

Antes de comenzar con la elaboración de páginas, debemos tener en cuenta dos aspectos fundamentales que influirán decisivamente en la forma en que nos enfrentamos a estas transcripciones. Esto es, la incorporación o no de reconocimiento óptico de caracteres.

- Para documentos sin OCR se creará una página desde cero, transcribiendo todos sus elementos, atendiendo además a los formatos del lenguaje wikitexto.
- Si el archivo dispone de una capa de texto con OCR, la actividad principal será la corrección de fallos, así como su formateo. Wikisource incorpora esta capa de texto de forma automática y la muestra en la pantalla de edición de páginas frente a su imagen digitalizada.

<sup>17</sup> Un wikitexto es un texto elaborado mediante un lenguaje de marcado especial para la creación de páginas wikis, en servidores que tengan instalado algún software para wikis. El lenguaje de marcado para wikis se denomina lenguaje wikitexto. Wikitexto. (2017). *Wikipedia, La enciclopedia libre*.

En el caso de que los archivos digitalizados no incorporen reconocimiento óptico de caracteres, desde Wikisource tenemos la posibilidad de hacer uso del *gadget* OCR. Al seleccionarlo dentro de la edición de una página, se envía una petición de reconocimiento de texto sobre la imagen digitalizada al servidor toolserver basado en Tesseract<sup>18</sup>, y al cabo de unos segundos aparece el texto leído. La corrección se realiza entonces sobre los resultados obtenidos con el OCR que pone a disposición Wikisource.

(véase Anexo 2).

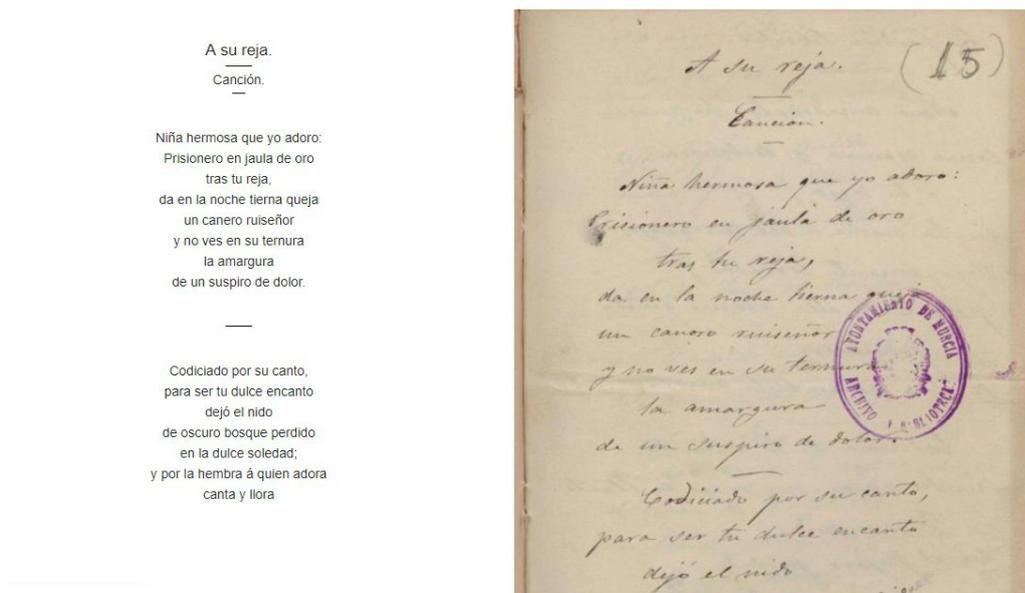


Figura 12. Resultado de transcripción (izquierda) de la imagen digitalizada (derecha).

Wikitexto:

```
<big>{{c|A su reja.}}</big>
```

```
{{línea|2em}}
```

```
{{c|Canción.}}
```

```
{{línea|1em}}
```

```
<br>
```

```
{{c|Niña hermosa que yo adoro:<br>Prisionero en jaula de oro<br>tras tu reja,<br>da en la noche tierna<br>queja<br>un canero ruiseñor<br>y no ves en su ternura<br>la amargura<br>de un suspiro de dolor.}}
```

```
<br>
```

```
{{línea|2em}}
```

```
<br>
```

```
{{c|Codiciado por su canto,<br>para ser tu dulce encanto<br>dejó el nido<br>de oscuro bosque perdido<br>en la dulce soledad;<br>y por la hembra á quien adora<br>canta y llora}}
```

<sup>18</sup> Tesseract es un motor de reconocimiento óptico de caracteres de software libre. En 2006 Tesseract fue considerado uno de los motores de OCR de código abierto más precisos, y su desarrollo ha sido patrocinado por Google desde ese mismo año. Tesseract (software). (2017). *Wikipedia, The Free Encyclopedia*.

Antes de guardar la página transcrita o corregida debemos indicar el nivel de revisión con los botones *checkbox* situados bajo la caja de edición. Hay cuatro niveles posibles:

- Sin texto: señala aquellas páginas sin texto, tapa, contratapa, etc. No es necesario corregir ni validar este tipo de páginas. Se presenta en color gris.
- Problemática: indica que existe algún tipo de problema con el texto. Esto será tratado en la página de discusión. Se presenta en color lila.
- No corregido: valor por defecto antes de la creación de páginas. Se presenta en color rojo.
- Corregido: este estado se selecciona cuando ha participado una sola persona en la transcripción o corrección. Se presenta en color amarillo.



Figura 13. Selección de estados de página.

Por último, cuando un usuario edita una página *corregida* por otro usuario, aparece una quinta opción de *validado*. Este paso asegura que más de una persona se encargue de corregir las transcripciones, reduciendo los posibles fallos humanos.

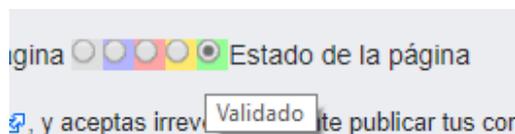


Figura 14. Validación de páginas colaborativa.

#### 6.2.2.2 Transclusión o publicación de textos

En último lugar, para que un texto esté disponible para todos debe estar publicado en el espacio de nombres *Principal*. La **transclusión** es el proceso que permite visualizar el contenido de una página de Wikisource en otra, de manera íntegra y sincronizada, para facilitar la edición y mantenimiento de los textos, además de su visualización. Consiste en traspasar el contenido de las páginas escaneadas del texto original al espacio de nombres *Página*: (en el paso anterior veíamos el proceso de creación, transcripción y corrección de páginas) y luego transcluirlas al *Principal* (donde se presenta el texto).

Para la transclusión de las páginas corregidas en una nueva página simplemente debemos crearla señalando las páginas de una obra que deben reunirse para formar el documento final.

<b>Autor</b>	Ricardo Sánchez Madrigal
<b>Título</b>	A su reja: canción
<b>Año</b>	1874
<b>Imprenta o editor</b>	
<b>Fuente</b>	Biblioteca Digital de la Ciudad de Murcia - MU-AM, 10-E-15(15) <a href="#">↗</a>
<b>Progreso</b>	Terminado — Todas las páginas validadas
<b>Páginas</b>	1 2 3 4 5 6 7 8 9 10 11 12 13

Figura 15. Páginas corregidas (5, 6 y 7) susceptibles de transcluirse en el documento final.

Por lo tanto, el código de esta nueva página no debe contener las transcripciones de cada una de las páginas, sino que reúne los *resultados* de éstas:

```
<pages index="A su reja.pdf" from=5 to=7/>
```

Desde la página final donde se muestra la transcripción completa del documento disponemos de enlaces a las páginas junto a su correspondiente imagen digitalizada. De esta forma podemos revisar que el texto que se ha transcrito es fiel a la digitalización. Los errores de transcripción pueden ser instantáneamente revisados y corregidos por cualquier usuario.

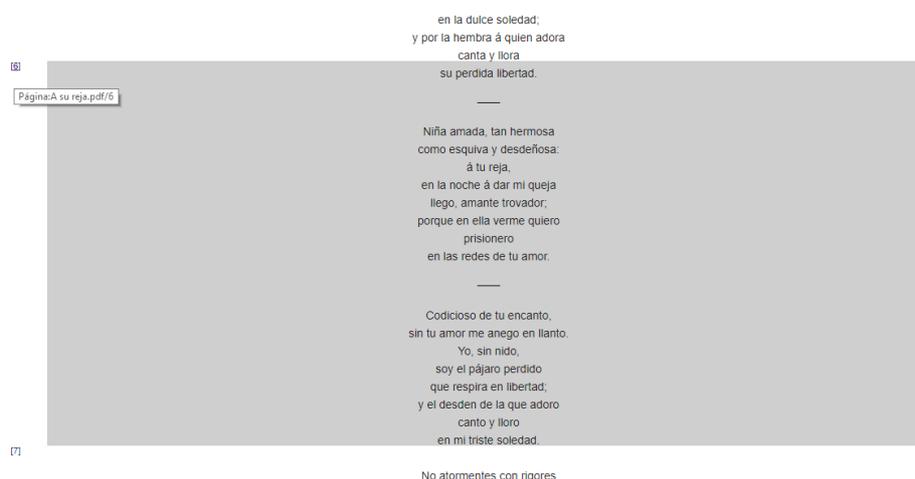


Figura 16. Obra "A su reja" con enlace a las páginas de transcripción.

Por último, para facilitar la navegación, y con un motivo estético, es recomendable añadir plantillas para portadas, encabezados, o índices.

### 6.2.2.3 Catalogación

Los documentos o textos contenidos en Wikisource se presentan lo más integros posibles, evitando enlaces y vínculos. Para la tarea de enlazar y vincular documentos se usan las **categorías**.

Los siguientes elementos deben categorizarse:

- **Obras literarias.** Al categorizar las obras en Wikisource estamos definiendo éstas, a la vez que las vinculamos con otras obras disponibles que comparten las mismas características. Existen cuatro categorías imprescindibles para los documentos:

1. **[[Categoría:ES-X]]**: esta categoría se refiere a la primera letra con la que empieza la obra. Wikisource presenta un índice alfabético de obras literarias.

Índice de obras literarias													
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z	

Figura 17. Listado alfabético de las obras literarias disponibles en Wikisource.

Para el poema "A su reja" de Ricardo Sánchez Madrigal categorizamos **[[Categoría:ES-A]]**, pasando el poema a formar parte de los 1631 documentos disponibles cuya primera letra es A.

2. **[[PXXXX]]**: esta categoría se refiere al año de publicación de la obra. Para una obra publicada en el año 1878, categorizamos **[[P1878]]**.

En casos en que se desconozca la fecha de publicación exacta de una obra o en el caso de obras que tengan fechas de publicación conflictivas esta categoría puede ser obviada. En el poema "A su reja" obviamos esta categoría al desconocerse su fecha exacta de publicación.



Figura 18. Listado de obras por fecha.

3. [[Categoría:Novelas]] o [[Categoría:Cuentos]], etc.: esta categoría se refiere al género literario de la obra. Éste puede ser novelas, cuentos, poesías, etc. Wikisource ofrece una lista completa de géneros en la Categoría:Literatura por género.

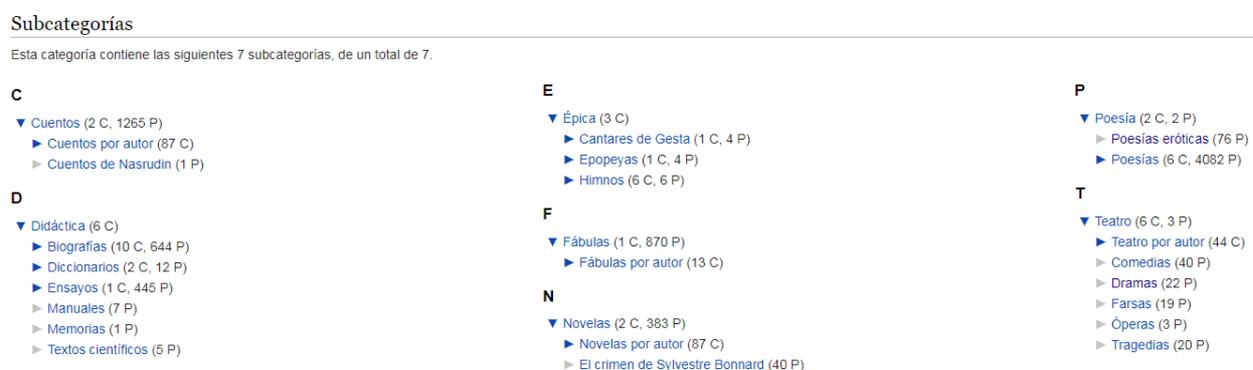


Figura 19. Subcategorías de literatura.

Wikisource establece un total de 7 géneros: cuentos, didáctica, épica, fábulas, novelas, poesía y teatro. Éstos a su vez se encuentran divididos en subgéneros.

Para el poema "A su reja" de Ricardo Sánchez Madrigal categorizamos [[Categoría:Poesías]].

4. [[Categoría:Poesías de XXX]] o [[Categoría:Novelas de XXX]], etc.: esta categoría se refiere al género literario (el mismo de la categoría anterior) relacionado con el autor de la obra.

La obra "A su reja" es un poema escrito por Ricardo Sánchez Madrigal, por lo tanto categorizamos [[Categoría:Poesías de Ricardo Sánchez Madrigal]].

Por lo tanto, para la obra "A su reja" de Ricardo Sánchez Madrigal el código necesario para su correcta categorización es el siguiente:

```

[[Categoría:ES-A]]
[[Categoría:Poesías]]
[[Categoría:Poesías de Ricardo Sánchez Madrigal]]

```

- **Páginas de autor.** Para la creación de páginas de autor, Wikisource establece una categorización automática.

Para esta categorización automática es necesario aplicar la plantilla {{Biocitas}}, que por defecto será la forma en que se presenten las fichas de autor.



Figura 20. Modelo de ficha de autores.

La finalidad de la plantilla `{{Biocitas}}` es estandarizar las fichas de autor. Esta ficha contiene información biográfica sobre el autor, información sobre derechos de autor, un enlace a la categoría en la que están agrupadas todas sus obras, enlaces a otros proyectos, etc.

Hasta el año 2016 la plantilla tenía los siguientes parámetros de cumplimentación manual:

```

{{Biocitas
|Texto=
|Obras={{PAGENAME}}
|Documentos={{PAGENAME}}
|Foto=none
|Wikipedia={{PAGENAME}}
|Wikiquote={{PAGENAME}}
|Wikinoticias={{PAGENAME}}
|Commons={{PAGENAME}}
|AñosMuerte=AAAA
}}

```

Actualmente estos datos se capturan desde la base de datos Wikidata automáticamente. Para la captura automática de datos es necesario que un autor exista como elemento en Wikidata.

En esta fase fue necesario crear en Wikidata el elemento "Ricardo Sánchez Madrigal" que recibe como nombre el código Q32860735. A partir de las declaraciones realizadas sobre el elemento, se establece su lugar y fecha de nacimiento, profesión, nacionalidad o fecha de fallecimiento.



Figura 21. Enlace de Wikidata hacia otros proyectos Wikimedia.

Debido a la captura automática de todos estos datos desde Wikidata, en Wikisource sólo es necesario establecer una ficha de autor con un pequeño texto sobre éste y su ordenación alfabética por apellido.

```

{{Biocitas
|Texto=""Ricardo Sánchez Madrigal""<br /> (22 de [[septiembre]] de 1844 - 12 de
[[junio]] de 1925)<br>Ingeniero de minas, escritor, periodista y poeta [[España |
español]].
|Ordenar=Sánchez Madrigal, Ricardo}}

```



Figura 22. Visualización final de categorías de Autor:Ricardo Sánchez Madrigal en su correspondiente ficha.

Con la catalogación de autores y sus obras se da por finalizado el proceso que comenzaba con la selección de obras, que eran depositadas en el repositorio Commons e integradas en Wikisource para su transcripción, corrección y publicación en forma de textos.

## 7. CONCLUSIONES

Se establecen las siguientes conclusiones para el presente trabajo:

- Se ha demostrado la utilidad de Wikisource como plataforma complementaria para la difusión de las colecciones digitales de bibliotecas.
- Esta utilidad se hace patente gracias a la visibilidad que otorga a las colecciones a través de su difusión, así como la posibilidad que se abre de reutilización de los materiales gracias a su integración en los diversos proyectos de la Fundación Wikimedia. Se valora además positivamente el posicionamiento favorable de los proyectos Wikimedia en Google.
- La plataforma Wikisource puede ser una forma de simplificar la parte tecnológica de proyectos de transcripción. Está disponible y no hay que hacer ninguna inversión ni mantenimiento en software o servidores propios.
- El segundo punto a favor en cuanto a costes es el ahorro en OCR: los usuarios realizan las labores de transcripción y corrección a la vez que se familiarizan e interactúan con las colecciones de las instituciones culturales.
- Los bibliotecarios que decidieran meterse en Wikisource deben adquirir formación directa previa en la plataforma, para tener desenvoltura en todo el proceso. Además implica manejarse también en Wikipedia, Commons y WikiData, por lo que el conjunto de destrezas se amplía.
- Por último, Wikisource permite establecer otro tipo de relación con los usuarios. Aunque el trabajo con Wikisource pueda parecer en cierto modo una replicación de fuentes y de trabajo, puede llegar a tener valor ya que, en internet, la pluralidad de versiones es habitual y puede verse como una forma de llegar a comunidades de usuarios distintas.
- La plataforma wiki también permite coordinar comunidades de usuarios a través de wikiproyectos o páginas de discusión.
- El trabajo realizado puede servir de base para la elaboración de un material para sesiones de formación práctica para bibliotecarios, de forma que sirva para reducir barreras de entrada y permitir tomar decisiones informadas.

A continuación se exponen las conclusiones a las que se ha llegado en cada una de las hipótesis planteadas:

**Hipótesis 1:** *Wikisource produce un impacto positivo en la visibilidad y difusión de las colecciones digitales complementario a las acciones llevadas a cabo desde las instituciones culturales, que se traduce en un coste y unos beneficios para éstas.*

**Conclusiones a la hipótesis 1:** Algunos esfuerzos no resultan en tráfico directo hacia los repositorios de bibliotecas digitales, pero gracias al uso de proyectos como Wikisource se crea conciencia de los recursos disponibles en las instituciones. Atraer a los usuarios potenciales a nuestros espacios no debería ser la primera ni la única solución: las bibliotecas deben acercarse al mundo que el usuario ya trabaja a diario y conoce bien. Es por todo esto que Wikisource se presenta como un recurso idóneo para las instituciones culturales debido a que son los usuarios los que realizan las tareas que éstas no pueden permitirse por falta de presupuesto o personal.

**Hipótesis 2:** *Wikisource permite lograr otro tipo de relación con los usuarios y grupos de interés que resultan en la consecución de objetivos de servicio al usuario.*

**Conclusiones a la hipótesis 2:** El actual contexto donde encontramos y nos relacionamos con los usuarios de información cultural está cambiando. Dentro de este cambio Wikisource ofrece nuevas herramientas para conectar con estos grupos, dando lugar a nuevos entornos para el intercambio de información. Para dar importancia al hecho de la unión del usuario y la información, iniciativas como la estrategia Europea 2020 resaltan la importancia de enriquecer y mejorar los datos a través de eventos relacionados con el usuario y la literatura. Anima, de igual manera, a que desde las propias bibliotecas, archivos y museos se trabaje con el usuario para aumentar la calidad de los datos y ampliar el compromiso público con las instituciones de una forma innovadora. La implicación de las instituciones en proyectos ya existentes como Wikisource permite conectar con comunidades de prácticas y de interés establecidas: que reúnen una identidad, afinidades y sobre todo unos objetivos comunes. Una vez se asiente la dinámica de trabajo y se consiga una masa crítica de usuarios a la que se pueda denominar *comunidad*, las instituciones pueden plantearse la creación de plataformas propias, desde Wikisource se complementan los esfuerzos.

## REFERENCIAS BIBLIOGRÁFICAS

- Alcalá, M. (2015). Crowdsourcing en las instituciones de la memoria: las transcripciones masivas. *BiD: textos universitaris de biblioteconomia i documentació*, 35. Recuperado de <http://bid.ub.edu/es/35/alcala.htm>
- Anderson, N. (s.f.). *Optical Character Recognition: IMPACT Briefing Paper*. Recuperado de [https://www.digitisation.eu/download/website-files/BP/OpticalCharacterRecognition-BriefingPaper\\_01.pdf](https://www.digitisation.eu/download/website-files/BP/OpticalCharacterRecognition-BriefingPaper_01.pdf)
- Anderson, N.; Muhlberger, G. y Antonacopoulos, A. (s.f.). *Optical Character Recognition: IMPACT Best Practice Guide*. Recuperado de [https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG\\_01.pdf](https://www.digitisation.eu/download/website-files/BPG/OpticalCharacterRecognition-IBPG_01.pdf)
- Andro, M. (2016). *Bibliothèques numériques et crowdsourcing: expérimentations autour de Numalire, projet de numérisation à la demande par crowdfunding* (Tesis doctoral. Université Paris 8, Francia). Recuperado de <http://www.theses.fr/2016PA080039>
- Arquero, R. y Cobo, S. (2014). Modelo de negocio y visión estratégica de Europeana. En L. Ramos y R. Arquero. (Coords.), *Europeana: la plataforma del patrimonio cultural europeo* (p. 59-88). Gijón: Trea.
- Ballesteros, S.; Morales, G. y Cedill, P. A. (2012). Los problemas de identificación de caracteres OCR para la recuperación de texto en el libro antiguo: un análisis de caso en el Fondo Antiguo de la Biblioteca Central, UNAM. *Biblioteca Universitaria*, 15(1), 25-34. Recuperado de <http://www.redalyc.org/articulo.oa?id=28528264003>
- Biblioteca Nacional de España. (2017a). *Autores en dominio público*. Recuperado de <http://www.bne.es/es/Servicios/InformacionBibliografica/AutoresDominioPublico/>
- Biblioteca Nacional de España. (2017b). *Autores en dominio público en 2017*. Recuperado de [http://www.bne.es/webdocs/Servicios/Informacion\\_bibliografica/autores-dominio-publico-2017.pdf](http://www.bne.es/webdocs/Servicios/Informacion_bibliografica/autores-dominio-publico-2017.pdf)
- Bocanegra, J. (2016). *Mejoras en el OCR Tesseract*. (Trabajo Fin de Grado. Universidad de Sevilla, Sevilla). Recuperado de <http://bibing.us.es/proyectos/abreproy/90978/fichero/Jes%C3%BAs+Bocanegra+-+Mejoras+en+el+OCR+Tesseract.pdf>

- Bremner, E. (2011). *Introduction to digitisation*. Recuperado de [https://www.digitisation.eu/download/website-files/BP/IntroductionToDigitisation\\_01.pdf](https://www.digitisation.eu/download/website-files/BP/IntroductionToDigitisation_01.pdf)
- Brumfield, B. (2015). *Collaborative Manuscript Transcription*. En Crowdsourcing Consortium (Ed.), *Engaging the Public: Best Practices for Crowdsourcing Across the Disciplines, May 6-8 2015* (pp. 93-95). University of Maryland. Recuperado de <http://www.crowdconsortium.org/wp-content/uploads/crowdconP.pdf>
- Carrera, C. (2012). *La biblioteca digital*. Gijón: Trea.
- Cobo, C. (2007). Intercreatividad y Web 2.0: la construcción de un cerebro digital planetario. En Cobo, C. y Pardo, H. (Coords), *Planeta Web 2.0. Inteligencia colectiva o medios fast food* (p. 43-59). Barcelona: Grup de Recerca d'Interaccions Digitals, Universitat de Vic. Recuperado de <http://www.planetaweb2.net/>
- Colmenero-Ruiz, M. J. (2015). Humanidades digitales y crowdsourcing: retos y oportunidades para bibliotecas y archivos. En *IV Seminario Hispano Brasileiro de Investigación en Información, Documentación y Sociedad: 24-26 de junio de 2015*. Marília: Universidade Estadual Paulista. Recuperado de [https://e-archivo.uc3m.es/bitstream/handle/10016/24023/Crowdsourcing\\_2015.pdf?sequence=4](https://e-archivo.uc3m.es/bitstream/handle/10016/24023/Crowdsourcing_2015.pdf?sequence=4)
- Comisión Europea. Comité des Sages on bringing Europe's cultural heritage online. (2011). *The New Renaissance: Report of the "Comité des Sages", Reflection Group on bringing Europe's cultural heritage online*. Brussels: European Commission. European Commission for the Digital Agenda.
- Cordal, J. (2016, 7 octubre). *La digitalización el patrimonio bibliográfico* [Vídeo]. Recuperado de [https://arcamm.uc3m.es/arcamm\\_3/item/show/f59df9db316d2c4ff817dcb8eed7f6b2](https://arcamm.uc3m.es/arcamm_3/item/show/f59df9db316d2c4ff817dcb8eed7f6b2)
- Dahlström, M., Hansson, J. y Kjellman, U. (2012). As We May Digitize: Institutions and Documents Reconfigured. *LIBER Quarterly*. 21(3-4), 455-474. DOI: <http://doi.org/10.18352/lq.8036>
- Dobrecky, L. P. (2016). Crowdsourcing en bibliotecas. *Revista de Bibliotecología y Ciencias de la Información*, n. 63, 71-77. Recuperado de <https://doi.org/10.5195/biblios.2016.297>
- Estellés-Arolas, E. y González-Ladrón-De-Guevara, F. (2012a). Clasificación de iniciativas

de crowdsourcing basada en tareas. *El profesional de la información*, 21(3), 283-291. Recuperado de <https://doi.org/10.3145/epi.2012.may.09>

- Estellés-Arolas, E. y González-Ladrón-De-Guevara, F. (2012b). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 32(2), 189-200. Recuperado de <https://riunet.upv.es/handle/10251/56904>
- Estellés-Arolas, E., Navarro-Giner, R., y González-Ladrón-de-Guevara, F. (2015). Crowdsourcing Fundamentals: Definition and Typology. En *Advances in Crowdsourcing* (p. 33-48). Springer International Publishing. Recuperado de [https://www.researchgate.net/publication/276327717\\_Crowdsourcing\\_Fundamentals\\_Definition\\_and\\_Typology](https://www.researchgate.net/publication/276327717_Crowdsourcing_Fundamentals_Definition_and_Typology)
- Europeana (2015). *Report on the results of the Wikimedia Taskforce*. Recuperado de [http://pro.europeana.eu/files/Europeana\\_Professional/Europeana\\_Network/europeana\\_wikimedia\\_taskforce\\_report\\_2015.pdf](http://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/europeana_wikimedia_taskforce_report_2015.pdf)
- Europeana 2020. (s.f.). Recuperado de <http://strategy2020.europeana.eu/update/>
- Europeana Foundation (2017). *Europeana Content Strategy: getting the right content to the right user at the right time*. Recuperado de [http://pro.europeana.eu/files/Europeana\\_Professional/Publications/Europeana%20Content%20Strategy.pdf](http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Content%20Strategy.pdf)
- Europeana Strategy 2015-2020. (s.f.). Recuperado de <http://strategy2020.europeana.eu/>
- García, I. (2016). La evolución de la escritura en la Corona de Castilla entre los siglos XIII y XVIII. En *De la Gótica al Whatsapp: taller de Historia del Archivo General* (p. 9-31). Murcia: Tres Fronteras.
- IFLA e ICA (2014). *Directrices para proyectos de digitalización de colecciones y fondos de dominio público, en particular para aquellos custodiados en bibliotecas y archivos*. Madrid: Ministerio de Cultura. Secretaría General Técnica.
- IFLA (2016). *Opportunities for Academic and Research Libraries and Wikipedia*. Recuperado de <http://2016.ifla.org/wp-content/uploads/sites/2/2016/08/112-IFLAWikipediaAcademicandResearchLibrariesDiscussioDRAFT.pdf>
- López, J. M. (1998). Normas españolas para la transcripción y edición de colecciones diplomáticas. *Espacio, Tiempo y Forma, Serie III, Hª Medieval*, t. 11, 285-306. Recuperado de <http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:ETFFF5A19F7->

- Lorente, D. (2016). *Estudio de situación de las colecciones y secciones de temas locales de las bibliotecas públicas de la Región de Murcia: resultados de una encuesta*. (Trabajo Fin de Grado. Universidad de Murcia, Región de Murcia). Recuperado de <http://hdl.handle.net/10201/52569>
- Martínez, J. M. (2016). Tecnoescritura: la escritura del tiempo presente. En *De la Gótica al Whatsapp: taller de Historia del Archivo General* (p. 85-114). Murcia: Tres Fronteras.
- Mendo, C. y Tejada, C. M. (2014). Europeana: un recorrido desde su nacimiento hasta nuestros días. En L. Ramos y R. Arquero. (Coords.), *Europeana: la plataforma del patrimonio cultural europeo* (p. 25-44). Gijón: Trea.
- Ministerio de Educación, Cultura y Deporte (2016). *Requisitos técnicos de los proyectos de digitalización de patrimonio bibliográfico y de prensa histórica de la SGCB*. Madrid: Subdirección General de Coordinación Bibliotecaria. Recuperado de <http://hdl.handle.net/10421/8981>
- National Library of Australia. (s.f.). *Optical Character Recognition (OCR) on Newspapers*. Recuperado de <https://www.nla.gov.au/content/ocr-overview>
- Neudecker, C. y Tzadok, A. (2010). User Collaboration for Improving Access to Historical Texts. *LIBER Quarterly*, 20(1), 119–128. Recuperado de <http://doi.org/10.18352/lq.7981>
- Pastor-Sánchez, J. y Saorín, T. (2011). Bancos de imágenes para proyectos enciclopédicos: el caso de Wikimedia Commons. *El profesional de la información*, 20(4), 424-431.
- Pereira, L.A. (2014). E se a biblioteca de Alexandria fosse um wiki?: relato de voluntariado. En *III Seminário De Pesquisa Da FESPSP*. São Paulo: Fundação Escola de Sociologia e Política de São Paulo. Recuperado de [http://www.fespsp.org.br/seminario2014/anais/GT6/3\\_E\\_SE\\_A%20BIBLIOTECA.pdf](http://www.fespsp.org.br/seminario2014/anais/GT6/3_E_SE_A%20BIBLIOTECA.pdf)
- Reagle, J. (2010). *Good faith collaboration: The culture of Wikipedia*. Cambridge (Mass): MIT Press.
- Real Decreto 224/2016, de 27 de mayo, por el que se desarrolla el régimen jurídico de las obras huérfanas, BOE núm. 141 § 5717 (2016). Recuperado de <https://boe.es/buscar/pdf/2016/BOE-A-2016-5717-consolidado.pdf>
- Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido

de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia, BOE núm. 97 (1996). Recuperado de <https://www.boe.es/buscar/pdf/1996/BOE-A-1996-8930-consolidado.pdf>

- Robledano, J. (2016, 22 marzo). La digitalización y acceso al patrimonio documental. El papel de las bibliotecas digitales interinstitucionales [Entrada blog]. Recuperado de <http://clip.sedic.es/article/la-digitalizacion-acceso-al-patrimonio-documental-papel-las-bibliotecas-digitales-interinstitucionales/>
- Rodríguez-Posada, Emilio-José y Saorín, T. (2012). Análisis de enlaces hacia bibliotecas y archivos digitales de patrimonio cultural desde Wikipedia en español y catalán. *BiD: textos universitaris de biblioteconomia i documentació*, n. 28.
- Rodríguez-Yunta, L. (2013). Humanidades digitales, ¿una mera etiqueta o un campo por el que deben apostar las ciencias de la documentación? *Anuario ThinkEPI*, v. 7, 37-43.
- Roig, A., Sánchez-Navarro, J. y Leibovitz, T. (2017). Multitudes creativas: el crowdsourcing como modelo para la producción audiovisual colectiva en el ámbito cinematográfico. *El profesional de la información*, 26(2), 238-248. Recuperado de <https://doi.org/10.3145/epi.2017.mar.10>
- Saorín, T. (2011). La difusión y creación de contenidos digitales locales en el ámbito cultural. T. Saorín y M. V. de Haro (Coords.), *Contenidos digitales locales: modelos institucionales y participativos* (p. 11-26). Madrid: Federación ANABAD.
- Saorín, T. (2012). *Wikipedia de la A a la W*. Barcelona: Editorial UOC.
- Saorín, T. y Rodríguez, E. J. (2012). Análisis de enlaces hacia bibliotecas y archivos digitales de patrimonio cultural desde Wikipedia en español y catalán. *BiD: textos universitaris de biblioteconomia i documentació*, núm. 28. DOI: <http://dx.doi.org/10.1344/105.000001816>
- Saorín, T. (2013). Iniciativas GLAM-Wiki: Wikipedia como oportunidad para instituciones culturales. *Anuario ThinkEPI*, vol. 7, 78-85.
- Saorín, T. (2017, 1 junio). Wikipedia en las instituciones de la memoria: nuevos territorios, nuevas costumbres [Vídeo]. Recuperado de [https://www.youtube.com/watch?v=1z\\_85LHV2Go](https://www.youtube.com/watch?v=1z_85LHV2Go)
- Segaran, T. (2007). *Programming Collective Intelligence*. California: O'Reilly Media.
- Serra, E. (2011). *Europeana y el Renacimiento digital europeo: la visión de la Comisión*

Europea. *Blok de BiD.* Recuperado de <http://www.ub.edu/blokdebid/es/content/europeana-y-el-renacimiento-digital-europeo-la-visi%C3%B3n-de-la-comisi%C3%B3n-europea>

- Serra, E. (2017, 12 mayo). A colaboración aberta nas institucións da memoria [Vídeo]. Recuperado de [https://youtu.be/fk8c59fQhJg?list=PLU-Ltp4IaHf3hFk3apIhN4vBe3Ztxl\\_S](https://youtu.be/fk8c59fQhJg?list=PLU-Ltp4IaHf3hFk3apIhN4vBe3Ztxl_S)
- Sundin, O. (2011). Janitors of knowledge: constructing knowledge in the everyday life of Wikipedia editors. *Journal of Documentation*, 67(5), 840-862.
- Tramullas, J. (2014). Wikipedia: de enciclopedia a objeto de investigación. *Anuario ThinkEpi*, vol. 8.

## ANEXOS

### ANEXO 1.

Sintaxis Wiki.

Formato	Significado
==Nivel 1==	Cabecera de primer nivel
===Nivel 2===	Cabecera de segundo nivel
====Nivel 3====	Cabecera de tercer nivel
"cursiva"	<i>cursiva</i>
""negrita""	<b>negrita</b>
*Entrada	*Líneas en listas no numeradas
#Entrada	#Listas enumeradas
-	Línea horizontal
[[Título de nueva página]]	Enlace interno
[[Título   Otro texto para el título]]	Enlace interno con texto alternativo
http://www.ejemplo.com	Enlace externo (automáticamente se convierte en un enlace)
[http://www.ejemplo.com/ www.ejemplo.com]	Enlace externo con texto alternativo
[[Imagen:fichero.jpg]]	Inserta una imagen en la página
[[Imagen:fichero.jpg thumb]]	Inserta un thumbnail en la página
{{Nombre}}	Integra dinámicamente el contenido de la página "Plantilla:Nombre"
{{subst:Nombre}}	Reemplaza el contenido de "Plantilla:Nombre" cuando almacena la página
{{:Nombre}}	Integra dinámicamente el contenido de la página "Nombre"
~~~~	Crea una firma para el registro de usuario, esto es un enlace Wiki a una página personal del usuario
~~~~~	Firma con la marca temporal

## ANEXO 2.

Comparación de resultados de OCR obtenidos en Biblioteca Digital de la Región de Murcia y Wikisource.

Resultados de OCR con página problemática.

Biblioteca Digital de la Región de Murcia	Wikisource
<p>XII y su labor fue premiada con aplausos, do mismo ( uc la del Sr. Peris. líl' Sr - Bayarri fué cstrepiosainente ovacionado to das las - cces que salió á escena.. línclintermediodelprimeroalsegundoa cto de flavina. Se estrenó el monólogo Paco Cayuela, original del laureado poeta D. Pedro o Jara, por el distinguido artista á quien en el mundo de los sportmans (y £ en el otro de la demás gente) se conoce con aquél nombre. Cayuela estuvo hecho uñ actorazo, nos hizo reir grn nde mer ñe ' á todos y fué tan aplaudido, que de seguro grabará en letras de oro en el lib ro de sus triunfos la fecha de ayer. Al final del n. onólogo tuvo que presentarse en es cena Jai; Carrillo, obligado por los insistentes apl lausos del público. • El popularísimo sportman recibió r.umerosos regalos de sus amigos y admiradores. l'' clicitamos al amigo Bojart por el triunfo de anoche y ojalá jíoda mos aplaudirle muy pronto o el día en que pisé la escena de nuestro teatro, terminados po. - completo sus estudios artísticos. Que se apronto y con tai buen resultado c o mo anoche. €"/ Correo de Xevarjte «Siitifecho p u e d e estar n u e s t r o jó-en paisano de ias cariñosas de m o s t r a c i o n e s V ^ de simpatía que anoche le tributó el público, llenando la mayor parte de las localidades del teatro y tributándole repetidos é insistentes aplausos durante la representación de Marina.</p>	<p>mr jHC ki der Sin'í'crí's. ' " ' EY'Sr-ílBúyarri fué {stréfiilvsammte ovación- a'ó todas las veces que salió á escena. En. cl intermedio d'c'ljrizn'm-o al'segundo ac- t'ó de Marina; sc' estrenó el monólogo Paco Cnyucln, original del jmarcado poeta D.4Pedro Jara; pOr cl distinguido artista j"quien en el mundo de los sportmam ('yficn el otro de ja deb 'más 'gehtc) ac'conocc con aquél'nombre. Cayuela qstuvthcho'un actorazo, nos hito reir grandcméx'flc'fi todos y fué tan 'aplaudidg, que de s'e'guro grabará en letras dc orden el l'- 'bru'dc 'sus triun'fos la fecha de ayer. A] final del mpnólogo mm que presentarse en escena jam Carrillo. obligado por los iasistentes aplau- áos del púbgi'có. ' f 'El populñísimo spommn recibió numerosas r'ggáksídc' bus amigos y admiradores. 4 'Félicitamos al amigo'Bojart por cl triunfo de 'nnoch'c y bjakí podamos aplaudircl muy promo 'cl (lia cd que pise Ala escena de nuestro ' (tati- o, utcrminzldu's puccampcln sus estudios artísticos. :Qige scá promo y con tail buen resultado como anoche. ' ' '  fibbr'fiíe firéhñindu 'coñ "aplausos; lo mismo  v el ¿'9ch ae ¿»mm  z z «jatisfecho putada esta: nuestro jóven paisano de Jas'cariñosas dbn'ostzrlcionesfgdc simpatía qm: anoche: le tributó el público, llcnmldo' la " maydr parte de fas' localidades del jentra thri- bnlándolc rqmtidos é insistentes' aplausos .du-</p>

Resultados de OCR con página que no presenta problemática.

Biblioteca Digital de la Región de Murcia	Wikisource
<p>XtII</p> <p>A su presentación en escena, fué saludado con un cariñoso aplauso Javier Bojart, el cual cantó poseído de visible emoción el número de</p> <ul style="list-style-type: none"> <li>• salida.</li> </ul> <p>Después fué serenándose y pudo lucir su voz de agradable timbre aunque no de gran extensión, y el buen gusto en la emisión de aquella.</p> <p>Fué aplaudidísimo en casi todos los números de la zarzuela y en los de la ópera que también cantó: y al final de algunos de aquellos se le obligó á salir á escena.</p> <p>El precioso dúo con la tiple en el acto segundo, fué seguramente uno de los números que mejor le resultaron y que con mayor gusto se le aplaudieron.</p> <p>Además de los aplausos del público, el señor Bojart recibió varios regalos, entre ellos un alfiler de oro y una cantidad en metálico del exalcalde D. Lorenzo Pausa, otros en metálico de las Sras. D." Antonia Borja, Marquesa de Rioflorido y D.* Soledad Stárico, una pitillera de D. Benito López Ruano y un cuadro de D. Julián Calvo Gavilá</p> <p>En la interpretación de la popular zarzuela de Arrieta estuvo muy bien la Sra. Naya y admirable el Sr. Bueso, cuyas hermosas facultades le valieron una nueva ovación.</p> <p>El acontecimiento de la noche puede decirse que lo constituyó la presentación del popular -y simpático sportman de Totana, Paco Cayuela, con el monólogo escrito exclusivamente para este objeto por nuestro querido compañero de redacción, el poeta Pedro Jara Carrillo.</p> <p>La representación de dicho monólogo cons-</p>	<p>XIII</p> <p>A su presentación en escena, fué saludado con un cariñoso aplauso Javier Bojart. el cual cantó poseído de visible emoción el número de salida.</p> <p>Después fué serenándose y pudo lucir su voz de agradable timbre aunque no de gran extensión, y el buen gusto en la emisión de aquella.</p> <p>Fué aplaudidísimo en casi todo: los números de la zarzuela y en los de la ópera que también cantó: 'y al final de algunos de aquellos se le obligó á salir a escena.</p> <p>El precioso duo con la tiple en el acto segundo, fué seguramente uno de los números Aque mejor le resultaron y que con mayor gusto se le aplaudieron.</p> <p>Además de los aplausos del público, el señor Bojart recibió varios regalos, entre ellos un alfiler de oro y una cantidad en metálico del exalcalde l). Lorenzo Pausa, otros en metálico de las Sras. D.' Antonia Borja, Marquesa de Ríofo4 rido y D.l Soledad Stárieo. una pitillera de D. Benito Lopez Ruano y un cuadro de D. Julian Calvo Gavilá. l</p> <p>En la interpretacion dela popular zarzuela de Arrieta estuvo muy bien la Sra. Naya y admirable el Sr. Buceo. cuyas hermosas facultades le valieron una nueva ovación.</p> <p>El acontecimiento de la noche puede decirse que lo constituyó la presentación del popular .y simpático sportm de Totana. Paco Cayuela. con el monólogo escrito exclusivamente para este objeto por nuestro querido comino de redacción, el poeta Pedro Jara'Carrillo.</p> <p>La representación de dicho mólogo co»-</p> <p>I</p>