Edición de contenidos en un entorno colaborativo: el caso de la Wikipedia en español

Editing content in a collaborative environment: the case of the Spanish Wikipedia

Ángel F. ZAZO RODRÍGUEZ, Carlos G. FIGUEROLA, José Luis ALONSO BERROCAL

Instituto Universitario de Estudios de la Ciencia y la Tecnología, Edificio I+D+i - Universidad de Salamanca
{angelzazo | figue | berrocal}@usal.es

Resumen

Se analizan las características y la actividad que los usuarios editores de la Wikipedia en español realizan en el proceso de creación de contenidos. Tras volcar los datos de los artículos enciclopédicos, se han analizado aspectos cuantitativos de los artículos, como su longitud, enlaces entrantes y salientes entre ellos, y categorías a las que pueden ser asignados. En algunos casos, esas características siguen patrones similares a los encontrados en estudios webmétricos. El sistema de categorías, pese a que funcionalmente está bien constituido, no se utiliza de manera adecuada por los usuarios editores, lo cual menoscaba una buena forma de acceso al conocimiento. En cuanto a las ediciones que realizan los usuarios, se han obtenido patrones de actividad relacionados con la creación de artículos, la revisión de contenidos, días de actividad, reversiones, vandalismo y país de origen. Una parte importante del funcionamiento de Wikipedia recae en unos pocos usuarios que supervisan los nuevos contenidos, ayudados por robots que facilitan el proceso. En general, la creación de contenidos se lleva a cabo por dos grupos diferentes de usuarios: pequeñas contribuciones individuales de una gran legión de usuarios, y un gran número de contribuciones que realizan un reducido grupo de usuarios muy activos. Se ha obtenido el país de origen de muchos usuarios, lo cual ha permitido contabilizar las contribuciones realizadas desde cada uno de ellos.

Palabras clave: Wikipedia. Edición colaborativa. Estudios de usuarios. Organización del conocimiento.

Abstract

This work uses the database backup dumps that collect content and history reviews of the encyclopaedic articles of Spanish Wikipedia since its creation, in order to characterize and understand the underlying activity of the editors in content creation. Some quantitative characteristics of articles are analyzed: length, assigned categories and in-links and out-links to other articles. Some characteristics have similar patterns to the ones found in webometric studies. The categories system, even though is functionally well built, is not used properly by the editors, which undermines the access to knowledge. We have also obtained patterns of the editors' activity related to article creation, content reviewing, activity days, reversions, vandalism, and authors' countries of origin. We have found that an important part of Wikipedia lies on a few number of users who oversee the new content, aided by robots that facilitate the process. In general, content creation is performed by two different kinds of users: small individual contributions of a great legion of users and a large number of contributions made by a small group of extremely active users. For many users we have obtained the origin country, which has allowed us to know the contributions procedence.

Keywords: Wikipedia. Collaborative edition. User studies. Knowledge organization.

1. Introducción

Wikipedia se autodefine como una enciclopedia libre, políglota y editada colaborativamente. En poco tiempo ha cosechado un enorme éxito, por lo que ha sido estudiada desde muy diferentes puntos de vista. Son muchos los artículos académicos y de investigación que la tienen como objeto de estudio, existiendo incluso portales web dedicados a recopilar la literatura gris que sobre ella se genera, como por ejemplo, Wikilit (wikilit.referata.com), WikiPapers (wikipapers.referata.com) o AcaWiki (akawiki.org). La propia Fundación Wikimedia dispone de servicios para la difusión de investigaciones, como el *Wikimedia Research Newsletter*. La Wikipedia ha sido muy estudiada desde el punto de vista de su contenido: amplitud, tamaño, evolución, calidad, actualidad, fiabilidad, etc. (Hu et al., 2007; Luyt et al., 2008; Magnus, 2009; Lewandowski y Spree, 2011). Otras veces se ha analizado como escenario colaborativo para la edición: participación, motivación, reputación, calidad de las contribuciones, redes de autoría, vandalismo, etc. (Wu et al. 2010; Leskovec et al. 2010). También como fuente de conocimiento para diversos entornos, como salud, educación, deporte, noticias, etc. (véase http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_as_an_academic_source). Como gran corpus documental de ac-

ceso abierto que es, existen también muchos trabajos que la utilizan como fuente de datos en campos diversos, como la recuperación de información, el procesamiento de lenguaje natural, la construcción de ontologías y tesauros, etc. (Gabrilovich y Markovitch, 2009; Müller y Gurevych, 2009; Medelyan et al., 2009). Una buena recopilación de artículos sobre Wikipedia puede encontrarse en (Mesgari et al., 2015).

La inmediatez que ha supuesto Internet ha cambiado la forma en la que las personan intercambian conocimiento. Wikipedia es un ejemplo claro de este cambio. Al igual que *L'Encyclopédie* de Diderot y D'alembert supuso una revolución para su tiempo, al pretender recoger todo el conocimiento humano en una obra accesible a gran cantidad de personas, Wikipedia supone un ejemplo claro de colaboración voluntaria, que es consultada a diario por millones de personas.

La importancia de los procesos de organización y representación del conocimiento como intermediarios entre las personas y las fuentes de información, teniendo presente que la necesidad humana de informarse es consustancial al individuo, es lo que nos ha motivado a estudiar la Wikipedia desde el punto de vista de la creación de los contenidos que los usuarios de manera voluntaria realizan en esta enciclopedia.

Nuestro objetivo al estudiar la Wikipedia ha sido analizar de manera sistemática el proceso de creación y revisión de los artículos enciclopédicos. Se analizarán desde diferentes puntos de vista, desde su creación, su historial de revisión, los tipos de usuarios que editan y su origen, el uso de categorías y enlaces entre artículos (verdaderos elementos para representar el contenido), los procesos de reversión y vandalismo, etc. Previamente es importante conocer cómo Wikipedia organiza los contenidos y cómo los usuarios pueden acceder a ellos.

2. Funcionamiento de Wikipedia

La Wikipedia en español es la cuarta por número de páginas (páginas de todo tipo), es la décima por número de artículos enciclopédicos, y es la segunda en ratio de número de visitas por hora (véase http://wikistats.wmflabs.org/).

Wikipedia se basa en cinco pilares:

1. Solamente se aceptan contenidos enciclopédicos de carácter general, no se aceptan investigaciones originales ni ensayos, ni opiniones personales.
2. Se busca siempre el punto de vista neutral. Para ello deben ofrecerse todos los puntos de vista posibles, incluyendo fuentes autorizadas verificables. Las controversias deben solventarse en la página de discusión que poseen todos los artículos.
3. Cualquier persona puede incorporar, corregir o utilizar la información. El contenido está bajo licencia *Creative Commons* BY-SA 3.0.
4. Se exige respeto mutuo entre usuarios.
5. No hay otras normas aparte de las indicadas. Para colaborar en el proyecto solo es necesario guiarse por el sentido común, respetando en todo momento al resto de personas que en él colaboran.

El elemento fundamental de la Wikipedia es el artículo enciclopédico. Es el que recoge el conocimiento que buscan los lectores cuando consultan la Wikipedia. Todo artículo enciclopédico posee cinco elementos: el título, el contenido del artículo, la página de discusión en la que los usuarios editores realizan comentarios sobre el contenido, el apartado de edición que permite cambiar el contenido, y el historial de revisiones previas que hasta la fecha ha tenido el artículo.

Al igual que otras enciclopedias, Wikipedia incluye entradas que redirigen a otros artículos enciclopédicos. Tal es el caso de muchas siglas, acrónimos, seudónimos de personas y personajes, variantes léxicas del español y regionalismos, transliteraciones de nombres en otros idiomas, plurales, etc. A veces se utilizan para facilitar la navegación a usuarios que no puedan introducir determinados caracteres, como tildes, eñes, signos de puntuación invertidos o diéresis (por ejemplo, *Espana* redirige a *España*).

En el contenido de los artículos enciclopédicos se pueden incluir enlaces a otros artículos que permitan ampliar la información sobre dichos contenidos. Asimismo, cada artículo debe estar etiquetado en una o varias categorías. Su objetivo es poder encontrar grupos de artículos relacionados y navegar por ellos, ayudando a los lectores a conocer qué artículos existen sobre un determinado tema. Las categorías tienen a su vez subcategorías más específicas y supercategorías más generales, permitiendo navegar de lo general a lo concreto y viceversa. Al igual que los artículos, las categorías pueden ser editadas libremente. Existen categorías ocultas de mantenimiento que no se muestran a los usuarios.

Tanto los enlaces entre artículos como las categorías constituyen elementos de representación que permiten acceder a otros artículos que guardan relación con los contenidos (Pehcevski et al., 2010). Es importante, por consiguiente, estudiar cómo son utilizados por los usuarios editores. Hemos analizado la distribución del

número de artículos por categorías y del número de categorías por artículo. También hemos analizado la distribución del número de enlaces entrantes y salientes por artículo. En ambos casos el objetivo es determinar qué patrones de actuación siguen los editores al utilizar categorías y enlaces como mecanismos de acceso a otros contenidos, importantes desde el punto de vista del acceso a la información.

El segundo elemento importante de Wikipedia es el usuario editor (Sepehri et al., 2012). Los artículos pueden ser editados prácticamente por cualquier usuario, a excepción de algunos artículos protegidos contra el vandalismo y las guerras de ediciones. Todas las ediciones quedan identificadas por el usuario que las realiza, bien con el nombre de los usuarios registrados, bien con la dirección IP del ordenador desde donde editan los no registrados. Wikipedia no ofrece datos de los usuarios registrados, salvo la que los propios usuarios pueden incluyen en sus páginas de usuario, si desean crearlas.

Hay varios tipos de usuarios, caracterizados por las diferentes acciones que pueden llevar a cabo. La página *http://es.wikipedia.org/wiki/Wikipedia:Tipos_de_usuarios* explica de manera detallada los diferentes tipos. Desde el punto de vista de la edición de los artículos, en nuestro análisis hemos utilizado cuatro categorías:

1. *Usuarios no registrados:* Son aquéllos que no disponen de cuenta o no inician sesión, se identifican por una dirección IP y pueden crear o editar páginas no protegidas.
2. *Usuarios registrados:* Aquéllos que inician sesión y pueden crear o editar páginas. Son los usuarios nuevos y los usuarios confirmados y autoconfirmados. Estos últimos son usuarios con una experiencia de al menos 50 ediciones, que pueden editar páginas semiprotegidas y trasladar o renombrar páginas.
3. *Usuarios administradores:* Hemos reunido bajo esta categoría a varios tipos de usuarios. En general son aquellos que pueden realizar acciones de administración de páginas o de usuarios, como eliminar o proteger páginas, editar páginas protegidas, verificar contenidos, revertir ediciones, bloquear o cambiar derechos a usuarios, etc.
4. *Usuarios bots:* Este último grupo lo forman programas automáticos (robots) que recorren los artículos de la Wikipedia realizando diferentes acciones, como revisiones ortográficas, verificación de enlaces externos e *interwikis*, asistir en la desambiguación, verificación y traslado de categorías, marcado o reversión automática del vandalismo, etc.

Existen otros elementos en la Wikipedia, como páginas sobre convenciones y políticas de uso, páginas de usuario, páginas de plantillas, páginas de categorías, páginas de anexos, y también *páginas de discusión* de los diferentes espacios de páginas. Aproximadamente la mitad de las páginas de la Wikipedia se corresponden con este tipo de páginas, en ellas se dirimen diversas cuestiones y sirven de medio de comunicación entre usuarios. En este sentido son especialmente importantes las páginas de discusión de los artículos enciclopédicos, donde se solventan las posibles controversias sobre su contenido y organización (Wu et al. 2011).

Cuando se crea una página nueva o se modifica el contenido de alguna existente se ponen en marcha una serie de mecanismos que tienen el propósito de garantizar, de acuerdo a los cinco pilares de la Wikipedia, tanto la calidad de los contenidos como la uniformidad en su presentación. Tales mecanismos podemos englobarlos dentro de procesos de organización del conocimiento (Fernandez-Molina y Guimarães, 2002).

Las páginas nuevas aparecen en la página especial de *PáginasNuevas*, con un indicador para determinar si la página ha sido verificada. Los artículos nuevos creados por usuarios que cuentan con el permiso de verificado o autoverificado aparecerán, por defecto, como verificados. Ello permite a los usuarios verificadores centrarse en aquellos artículos donde el riesgo de violación de las políticas es mayor. Los usuarios verificadores pueden dar paso a las nuevas páginas, marcarlas para su borrado, o asignarles una de las plantillas de seguimiento.

Cuando se modifica el contenido de un artículo los cambios aparecen en la página especial de *CambiosRecientes*. Cualquier usuario de Wikipedia puede patrullar los cambios recientes para detectar ediciones dañinas, esto es, ediciones que no siguen las reglas de la Wikipedia sobre el contenido, a menudo errores sin malicia de usuarios nuevos; otras veces son acciones deliberadas, entre las que se incluyen las ediciones vandálicas y las guerras de ediciones (reversiones mutuas entre uno o varios usuarios motivadas por el contenido del artículo).

Los usuarios reversores (entre los que se incluyen varios *bots*) son aquellos que pueden revertir ediciones rápidamente a una versión previa cuando detectan vandalismo o ediciones que no siguen los cinco pilares. Una gran parte de las ediciones de los administradores y de los bots tienen que ver con las acciones de verificación y reversión de ediciones. Son de especial interés las reversiones motivadas por vandalismo, pues es habitual su análisis para determinar la cali-

dad de las contribuciones de los usuarios editores (Adler et al., 2011b).

En nuestro estudio hemos diferenciado las actuaciones que cada uno de los tipos de usuarios ha realizado en cuanto a la creación y a la revisión de los artículos, incluida la reversión de ediciones; el objetivo es conocer en detalle sus características. También hemos analizado la cantidad de información que los diferentes tipos de usuarios ha generado o destruido a lo largo del tiempo.

Por último, dado que la Wikipedia en español se nutre de editores procedentes de muchos países hispanohablantes, hemos analizado el país de origen de los mismos.

3. Metodología

Wikipedia funciona gracias al software libre MediaWiki. Todas las acciones que se realizan en Wikipedia quedan registradas en bases de datos. Un aspecto destacable desde el punto de vista de la investigación, es que la Fundación Wikimedia mantiene el historial de todas las acciones que se realizan en sus proyectos, y este historial, a excepción de los datos personales de los usuarios registrados, es público.

El acceso a los datos se puede realizar en tiempo real utilizando la interfaz de aplicaciones de MediaWiki o a las herramientas *Wikimedia Tool Labs* (http://wikitech.wikimedia.org). Datos generales se ofrecen en las páginas de estadísticas que poseen todos los proyectos, que se recopilan en http://stats.wikimedia.org/. Otra forma es accediendo al volcado completo que se realiza de todos los proyectos de forma regular en http://dumps.wikimedia.org. Los datos suelen estar comprimidos en ficheros XML y SQL.

En nuestro estudio hemos utilizado el volcado de datos de la Wikipedia en español de fecha 5 de enero de 2015. El volumen de datos tratados ha sido de 1 TB de información. Los datos fueron procesados con diferentes programas desarrollados *ad hoc* por los autores, para, en primer lugar, extraer la información y almacenarla en el gestor de bases de datos MariaDB (en parte utilizando los volcados SQL), y, en segundo lugar, para obtener los resultados que se presentan más adelante.

De todas las páginas de la Wikipedia en español hemos limitado nuestro análisis a las páginas que son artículos enciclopédicos, dado que son el objetivo principal de las personas que visitan Wikipedia. De hecho, la mayoría de estudios sobre Wikipedia ponen el punto de mira en los artículos enciclopédicos y en las acciones que realizan los usuarios editores sobre ellos (Kimmons, 2011; Yasseri y Kertész, 2013). Un aspecto a tener en cuenta es que hemos unificado los datos de todas las redirecciones de las entradas existentes al mismo artículo enciclopédico. Eso ha supuesto reducir el número de artículos de 2,7 a 1,1 millones.

Las características que hemos analizado de los artículos enciclopédicos son las siguientes: número de artículos, longitud media y distribución de artículos por longitud, número de categorías, distribución del número de categorías por artículo, distribución del número de artículos por categoría, número de enlaces entrantes y salientes, y las distribuciones asociadas por artículo. Varias de estas medidas se corresponden con medidas típicas de estudios cibermétricos y webmétricos (Berrocal et al., 2003; Stuart, 2013).

Los usuarios editores de la Wikipedia pueden crear artículos nuevos o modificar el contenido de los ya existentes. Hemos analizado ambos procesos, teniendo en cuenta los distintos tipos de usuarios. Los aspectos que hemos analizado son los siguientes: creación de artículos, número de artículos editados, distribución de usuarios editores considerando periodos de actividad y número de artículos editados, y cantidad de información creada o destruida.

Es importante señalar que hemos unificado todas las revisiones que un mismo editor ha realizado de manera sucesiva sobre el mismo artículo, utilizando para ello también un programa desarrollado por los autores. Por un lado, es bien sabido que una medida de prestigio de los editores es el número de ediciones realizadas (Adler et al., 2011a; Francke y Sundin, 2010). Una sencilla artimaña para aumentar ese número es realizar pequeños cambios sucesivos del mismo artículo. Por otro lado, muchos usuarios a menudo van incorporando información poco a poco en cada revisión, comprobando en cada paso el resultado del artículo. Con ambas unificaciones, el número de revisiones de artículos enciclopédicos se reduce significativamente de 55 a 37 millones.

Para determinar el país de origen de los editores hemos utilizado, por una parte, la geolocalización IP de los usuarios no registrados (se ha utilizado un programa desarrollado por los autores que utiliza las bases de datos de MaxMind, www.maxmind.com), y por otra, la información de aquellos usuarios registrados que lo indican es sus páginas de usuario, al asignarse categorías del tipo “Wikipedistas_de_España”, “Wikipedistas_de_Argentina”, etc. Este proceso se complica pues muchos editores se asignan categorías muy locales (por ejemplo, “Wikipedistas_de_La_Serena”), y ha sido necesario revi-

sarlas manualmente. Para usuarios muy activos se ha procedido también a revisar manualmente su página de usuario para determinar el país de origen. Los bots no tienen país de origen.

En cuanto a las reversiones, en nuestro estudio hemos detectado solamente las reversiones vandálicas, utilizando para ello información sobre el tipo de usuario que realiza la reversión y los comentarios que aparecen en el campo correspondiente de la base de datos.

4. Resultados y comentarios

Este apartado está organizado de acuerdo a los aspectos de interés que se han indicado previamente, para caracterizar los artículos enciclopédicos de la Wikipedia y los usuarios editores que los crean o los revisan. El primer apartado presenta datos cuantitativos de la Wikipedia en español a la fecha del volcado de datos. El segundo muestra resultados sobre la actividad de los diferentes tipos de usuarios editores, en lo referente a la creación de artículos, a la frecuencia de actuación y a la cantidad de información creada o destruida. El tercer apartado se centra en las revisiones de los artículos, ofreciendo resultados sobre número de artículos editados por editores humanos. También se incluyen en este apartado resultados relativos a las reversiones originadas por vandalismo. El último apartado ofrece resultados sobre el origen de los editores de la Wikipedia en español.

4.1. Artículos enciclopédicos

La Tabla I muestra el número de usuarios, el número de artículos creados y el número de revisiones (unificadas) por tipo de usuario, y el total, desde que la Wikipedia en español inicia su andadura en mayo de 2001 hasta la fecha del volcado de datos.

Tipos de usuarios-editores	*Nº de usuarios*	*Nº de art. creados*	*Nº de ediciones*
Registrados	557.970	422.874	7.980.795
Administradores	856	535.846	7.252.615
Bots	424	4.217	10.539.347
No registrados	5.793.501	148.863	11.179.884
TOTAL	6.352.751	1.111.800	36.952.641

Tabla I. Número y actividad de los usuarios de la Wikipedia en español

El número de artículos enciclopédicos es 1.111.800. Su longitud media es de 5.304 bytes, con una alta desviación de 9.616, lo cual indica una alta variabilidad. La distribución de frecuencias de artículos por longitud puede verse en el Gráfico I. A partir de los 10 KB de longitud la línea de tendencia sigue una ley de potencias inversa (ley de Zipf) con parámetro 0,5. Este parámetro es menor que el obtenido para estudios webmétricos (Baeza-Yates et al., 2005; Baeza-Yates et al., 2006), que está en torno a 2,7. Los dos artículos más largos son *Alsacia en 1789* e *Historia de Euskaltel-Euskadi*, con 800 KB y 736 KB respectivamente.

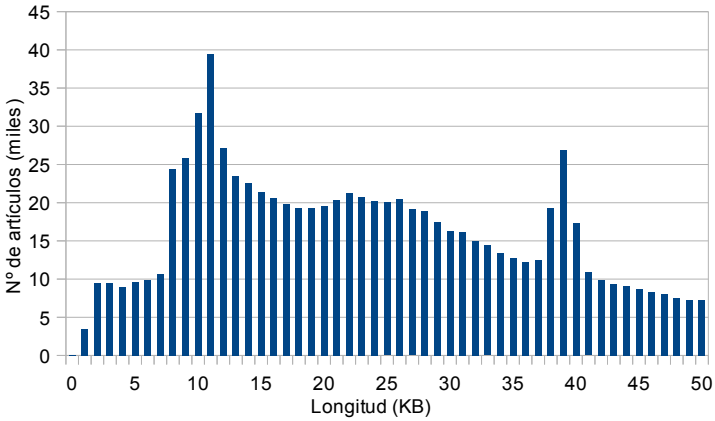

Gráfico I. Distribución de frecuencias de los artículos por longitud

El número de categorías no ocultas diferentes asignadas a los artículos es 227.152. La distribución de frecuencias de artículos por número de categorías puede verse en el Gráfico II. Hay unos pocos artículos sin categoría asignada. Se puede ver que la mayoría de ellos se etiquetan con una, dos o tres categorías. Los dos artículos asignados a un mayor número de categorías son *José María Pemán* y *Mario Vargas Llosa*, con 47 y 46 categorías, respectivamente.

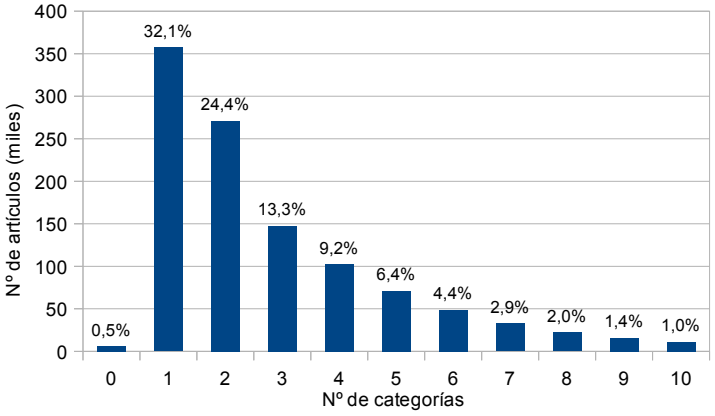

Gráfico II. Distribución de los artículos por categoría

Por otra parte, la inversa, esto es, la distribución de frecuencias de categorías por número de artículos se puede ver en el Gráfico III. Se aprecia que hay un 25% de categorías (57.156) que solamente aparecen en un artículo. Es decir, la asignación de categorías no siempre cumple el objetivo de que sirvan de navegación entre artículos. De acuerdo con este resultado podemos indicar que en la Wikipedia no se hace un uso adecuado de las categorías como elementos principales de representación del contenido para la obtención de nuevo conocimiento.

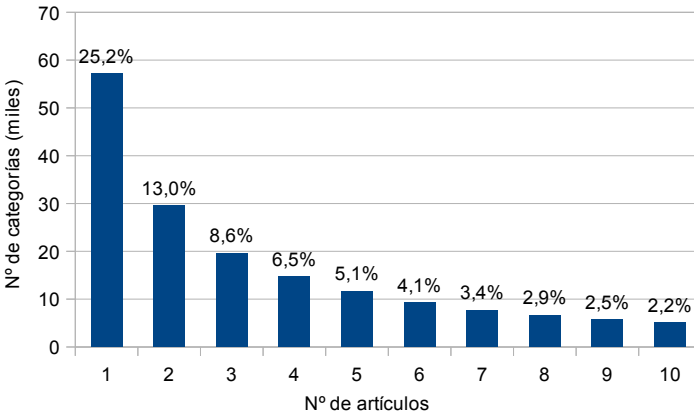

Gráfico III. Distribución de categorías por artículo

Se destaca que la categoría más frecuente es *Wikipedia:Desambiguación* con 46.243 artículos. Se trata de páginas con enlaces a artículos de títulos susceptibles se ambigüedad, en las que se ofrece una breve descripción de sus temáticas.

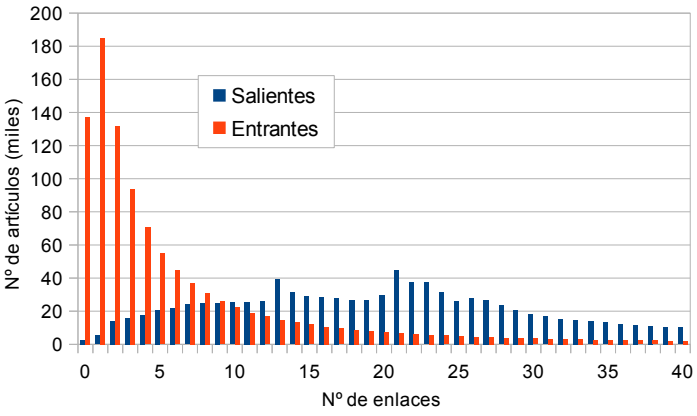

Gráfico IV. Distribución de frecuencias de enlaces entrantes y salientes por artículo

La distribución de frecuencias de enlaces entrantes y salientes, que permiten navegar directamente entre artículos, se muestra en el Gráfico IV. El número total de enlaces es de 33 millones. La distribución del número de enlaces entrantes es muy diferente a la de enlaces salientes. La distribución de enlaces entrantes tiene una distribución típica de larga cola o ley de potencias inversa, con parámetro 1,5. Este valor es ligeramente menor que para estudios cibermétricos sobre la Web (Baeza-Yates et al., 2005; Baeza-yates et al., 2006), que suelen estar en torno a 2,1. Es de señalar que hay un 12% de artículos que no reciben enlaces. Los dos artículos más citados son *Coordenadas geográficas* y *Estados Unidos*, con 235.971 y 197.008 enlaces entrantes, respectivamente.

La distribución del número de enlaces salientes, por el contrario, no sigue la típica distribución de potencias que sí se aprecia en estudios cibermétricos. Los usuarios editores suelen incorporar un elevado número de enlaces a otros artículos, con frecuencia a artículos de temática general o artículos relacionados solo tangencialmente con el contenido que están editando. Hemos comprobado que hay una correlación positiva de 0,76 entre la longitud de los artículos y el número de enlaces salientes a otros artículos. Los dos artículos que más enlaces salientes poseen son *Historia del Arte* y *Edad Contemporánea*, con 4.246 y 2.308 enlaces, respectivamente. Son dos artículos muy largos también.

4.2. Usuarios editores

El número total de usuarios aparecía en la Tabla I. De los 3,6 millones de usuarios registrados que ofrecen las estadísticas oficiales, en la práctica solamente han editado artículos 557.970. El número de bots es bastante elevado si lo comparamos con el de administradores. No obstante, la mayoría de ellos ha ejercido su labor durante unos pocos días y con muy pocas ediciones: suelen ser bots en pruebas (92 bots han realizado menos de 100 ediciones cada uno, estando activos 1 o 2 días). Por el contrario existen bots que están activos muchos días y realizan cientos de miles de ediciones: hemos comprobado que 31 bots han realizado más de 7,2 millones de ediciones, esto es, el 70% de todas las ediciones de los bots, el 20% de todas las ediciones (unificadas) de la Wikipedia. Si consideramos ediciones totales, sin unificar, el porcentaje de ediciones de todos los bots es del 19%. Este dato es específico para la Wikipedia en español, pues Wikipedias en otros idiomas ofrecen valores distintos (Mesgari et al., 2015).

En la misma tabla aparecen el número de artículos nuevos que cada uno de los tipos de usuarios ha creado, y las ediciones globales que ha realizado. Se destaca que el número de artículos que crean unos pocos administradores

es muy elevado; analizando más en profundidad los datos se comprueba que tan solo 28 usuarios (25 administradores y 3 usuarios registrados) han creado el 25% de los artículos. No obstante, no hay correlación significativa entre los usuarios que crean los artículos y los usuarios que los editan posteriormente. Estos resultados son similares para ediciones de Wikipedias de otros idiomas (Yasseri y Kertész, 2013; Gandica et al., 2015).

El Gráfico V muestra la evolución mensual en la creación de artículos para los diferentes tipos de usuarios. Desde el año 2006 hasta mediados del 2011 el ritmo de creación de artículos fue de unos 10.000 al mes. A partir de ese momento el ritmo ha sido de unos 7500 cada mes, siendo menor en el último año. Se aprecian meses de gran actividad: se debe a unos pocos editores que crean en muy poco tiempo una gran cantidad de artículos de la misma temática (astros, asteroides, plantas con ficha de taxón, coleópteros, arbustos, municipios, etc.). El número de artículos creados por usuarios administradores está muy por encima del resto. Vemos también que apenas existen artículos creados por bots, al contrario de lo que sucede en Wikipedias de otros idiomas, como la inglesa o la sueca (Okoli et al. 2012).

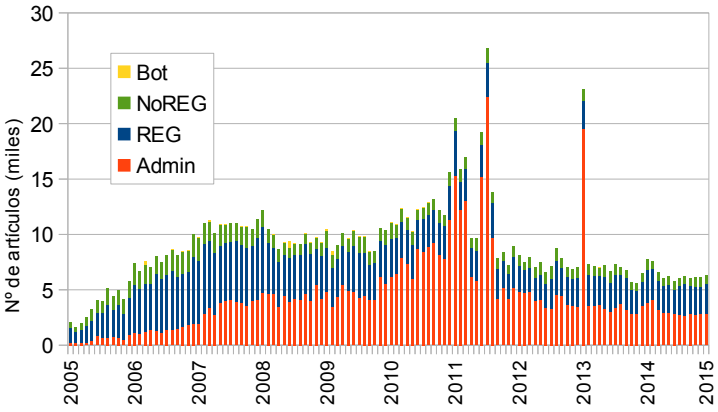

Gráfico V. Creación de artículos al mes

En el Gráfico VI se muestra la distribución de frecuencias relativas de editores por días de edición. La mayoría de usuarios anónimos y registrados están activos solamente un día, y editan un único artículo (Gráfico VII). Los usuarios administradores son los que más días dedican a labores de edición, y los que realizan más ediciones sobre un mayor número de páginas, excluidos, como era de esperar, los bots. El editor que más revisiones ha realizado es Diegusjaimes (156 mil), pese a dejar de editar artículos el 31-12-2012.

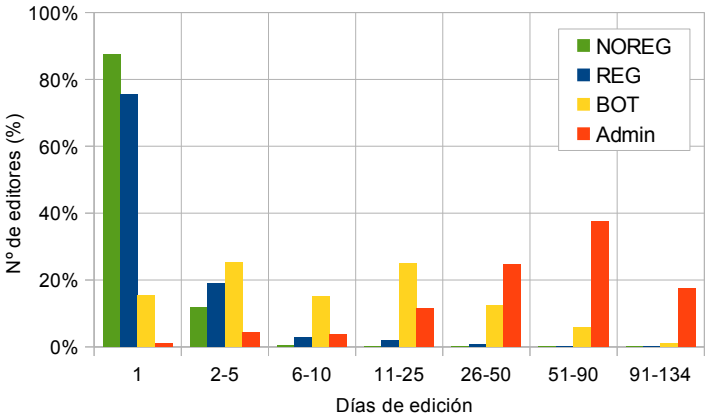

Gráfico VI. Distribución de frecuencias relativas de editores por días de edición

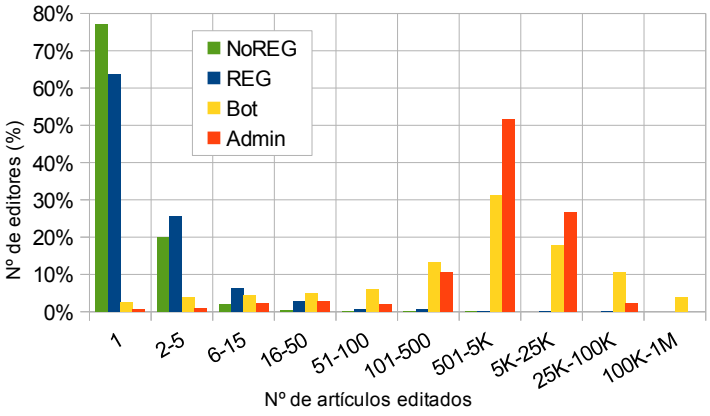

Gráfico VII. Distribución de frecuencias relativas de editores por número de artículos editados

Estas dos gráficas dan información muy valiosa sobre el protagonismo que tienen los diferentes tipos de usuarios en la Wikipedia. Podemos decir que existe un gran número de usuarios que realiza unas pocas contribuciones a la Wikipedia, y que una gran parte de las ediciones las realizan unos pocos usuarios administradores. Ello es general para todas las Wikipedias (Mesgari et al., 2015).

Hemos analizado también cuál es la actitud de los usuarios respecto de la creación de contenidos, teniendo en cuenta el número de bytes creados o destruidos en cada edición. En el Gráfico VIII se puede ver que desde el año 2007 se incorporan cada mes casi 50 MB nuevos a la Wikipedia. Hemos utilizado bytes y no palabras, como en estudios más completos (Adler et al., 2011a), porque hemos comprobado manualmente con varios cientos de artículos que los resultados son similares. La extraordinaria disminución de marzo de 2013 se debió al cambio de los enlaces *interwikis* entre Wikipedias de

diferentes idiomas al formato de WikiData, proceso que fue realizado por bots.

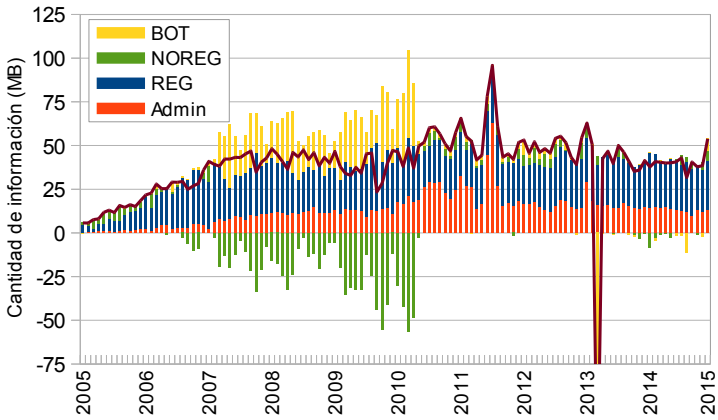

Gráfico VIII. Cantidad de información creada al mes

Se destaca también en el Gráfico VIII una actividad bastante negativa entre los años 2007 y 2010 de los usuarios no registrados. Este periodo coincidió con la publicación de la obra de Soliman y Gourdain (2008) que supuso una crítica tenaz contra Wikipedia, y muchos usuarios no registrados llevaron demasiado lejos algunas de sus propuestas de introducir pequeños cambios para determinar cuán rápidamente eran revertidos, realizando completos blanqueos de página. Podemos apreciar que en ese periodo de tiempo la actividad del resto de usuarios, especialmente de los bots, va casi pareja, pero en sentido contrario, a la de los usuarios no registrados, con lo que el volumen de información siguió incrementándose. Precisamente fue en este periodo cuando más se desarrollaron los bots antivandalismo.

4.3. Revisión de artículos

El Gráfico IX muestra el número de artículos que se revisan al mes una o más veces por editores humanos, no por bots, después de haber sido creados. Es un número bastante uniforme desde 2009, unos 100 mil al mes. Esto da idea de que el número de editores efectivos de Wikipedia no ha cambiado sustancialmente desde entonces. El número de revisiones sigue una tónica similar, aproximadamente desde 2009 cada mes se realizan unas 250 mil revisiones por editores humanos.

En el Gráfico X se muestra la distribución de frecuencias de artículos por número de editores humanos diferentes. Un 21% de los artículos (233.652) ha sido editado solamente por un editor; es decir, fue creado y no se ha vuelto a editar por nadie más. La mitad de los artículos han sido editados por cuatro o menos editores.

Los artículos más editados son, como cabía esperar, artículos muy antiguos, en los que han participado muchos editores, y que además han sufrido muchas reversiones. Los tres artículos más editados son *Venezuela, Club América* e *Idioma español*, creados todos ellos entre 2002 y 2005, con más de 6.200 revisiones cada uno de ellos, y con un porcentaje de reversiones en torno al 28% de todas sus revisiones.

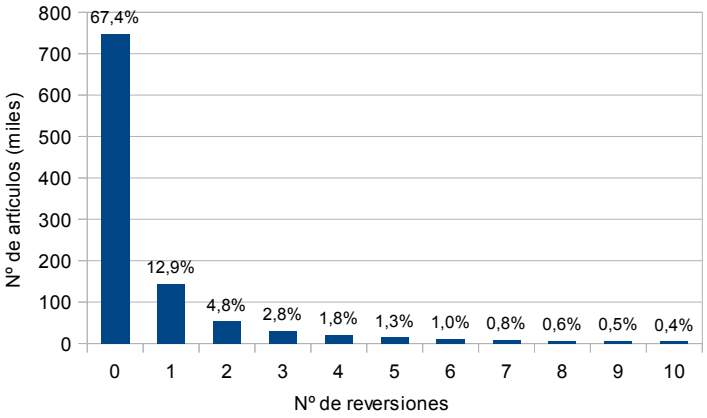

Gráfico IX. Artículos editados al mes por usuarios humanos, no por bots

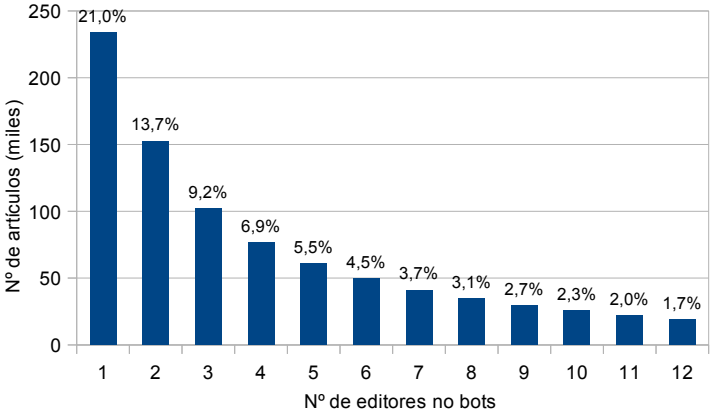

Gráfico X. Distribución de frecuencias de artículos por número de editores humanos diferentes

En el Gráfico XI se muestra la distribución de frecuencias de artículos por número de reversiones. El 67% de los artículos no se ha revertido nunca, y el 13% lo ha sido solamente una vez. Ello, junto con los resultados del gráfico anterior, nos indica que hay una buena disposición por parte de la mayoría de usuarios a la hora de incorporar contenidos a la Wikipedia. Los artículos más revertidos son *Club América*, *Dragon Ball*, *Baloncesto* y *Venezuela*. Son, como ya se indicó previamente, artículos muy antiguos.

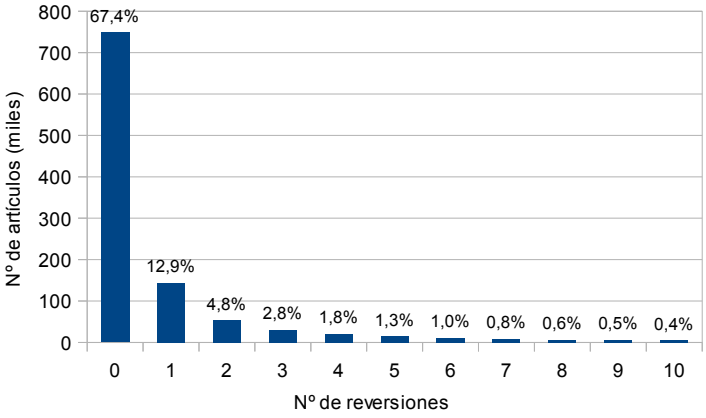

Gráfico XI. Distribución de frecuencias de artículos por número de reversiones

Los usuarios con atributos de reversor (administradores y algunos bots) son aquéllos que expresamente patrullan páginas para revertir vandalismos. Muchos usuarios registrados, sobre todo los más activos, también dedican mucho esfuerzo en esta labor. El número total de ediciones vandálicas revertidas por todos estos usuarios ha sido de 3,5 millones; es decir, que al menos otro número igual o mayor ha sido fruto del vandalismo. Al menos 3,5 millones de ediciones de un total de 19,2 millones de ediciones de usuarios de los tipos *registrado* y *no registrado* (los bots y los administradores no realizan ediciones vandálicas) son actos de vandalismo, lo cual supone un 18,3% de dichas ediciones. Es un porcentaje elevado, si bien, hemos comprobado que proceden de un número relativamente pequeño de usuarios. Los artículos más vandalizados son *Leyes de Newton, Romanticismo, Sistema operativo* y *Derechos humanos*.

Los reversones humanos se ayudan de bots que marcan posibles ediciones vandálicas, a menudo teniendo en cuenta usuarios (o direcciones IP) con historiales de acciones vandálicas previas. Algunos autores (Geiger y Ribes, 2010; Priedhorsky et al., 2007) indican que estas herramientas permiten que "voluntarios medios", que quizás tengan poco conocimiento del contenido del artículo en cuestión, puedan revertir ediciones vandálicas. Dicho proceso de revisión está en marcado contraste con las formas más tradicionales de producción de conocimiento profesional y académico de los expertos, que están en condiciones de contribuir a la Wikipedia en virtud de su conocimiento de un dominio determinado.

4.4. Origen de los editores

No debemos olvidar que el español es una de las lenguas más extendidas del mundo, y son muchas las contribuciones que se realizan desde muy diversos países a la Wikipedia en español. El Gráfico XII muestra el número de editores, creadores y ediciones realizadas en función de los países de origen de los editores para los que hemos podido obtener esa información (se indica el total en el gráfico). Las contribuciones latinoamericanas son muy importantes, por ello es frecuente encontrar variaciones dialectales, lo que en algunos casos ha llevado a algunas controversias.

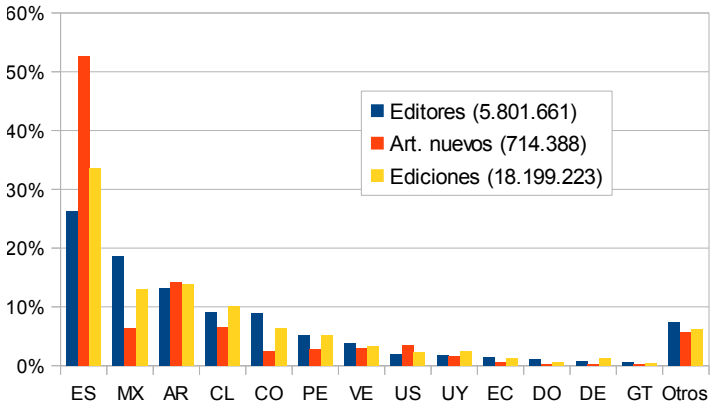

Gráfico XII. Origen de los editores por país (no bots)

5. Conclusiones

En este artículo hemos presentado el proceso de edición de la Wikipedia. Hemos estudiado los patrones de actuación que siguen los usuarios en sus contribuciones a esta herramienta de edición colaborativa. En cuanto a las características de los artículos enciclopédicos, hemos encontrado que para la longitud y los enlaces entrantes desde otros artículos se siguen distribuciones parecidas a las encontradas para el Web, si bien, con valores diferentes para los parámetros característicos. Para el caso de los enlaces salientes no existe tal analogía: ello se debe a que los usuarios editores suelen incluir gran cantidad de referencias a otros artículos enciclopédicos de temática general o relacionadas solo tangencialmente con el contenido que están editando.

En relación con el sistema de categorías que utiliza Wikipedia, cuyo objetivo es encontrar y navegar por grupos de artículos relacionados, hemos visto que no se utiliza adecuadamente como mecanismo de acceso a otros contenidos, tan importante en los sistemas de organización de la información. Es una de las deficiencias

más importantes que hemos encontrado, a pesar de que Wikipedia proporciona una buena infraestructura tecnológica al respecto y existen páginas con recomendaciones sobre el uso de categorías. Pensamos, por ello, que esta deficiencia se debe a una falta de sensibilidad o de formación de los usuarios editores a la hora de asignar categorías a los artículos.

Entre las conclusiones más importantes podemos destacar que la creación de contenidos se lleva a cabo por dos grupos diferentes de usuarios. Por un lado, pequeñas contribuciones individuales de una amplia legión de usuarios, en general usuarios registrados. Por otro lado, una gran cantidad de contribuciones que realizan un pequeño grupo de usuarios extraordinariamente activos, que poseen además atributos para la administración. La cantidad de información que se ha ido incorporando a la Wikipedia ha sido bastante uniforme en los últimos años, lo cual corrobora esta división de usuarios. La mitad de los artículos enciclopédicos han sido editados por cuatro o menos editores diferentes.

Destacamos también que una parte significativa del funcionamiento de Wikipedia recae en unos pocos usuarios que supervisan los contenidos nuevos que se van incorporando por usuarios anónimos o poco frecuentes, ayudados por bots que facilitan el proceso. Este resultado es general para Wikipedias de otros idiomas (Kimmons, 2011). Hay un gran número de revisiones que se realizan automáticamente por bots en busca de errores ortográficos y de formato, verificación de categorías y enlaces externos, etc., que ayudan al mantenimiento de la Wikipedia.

Una parte importante de las acciones de los usuarios administradores y de los usuarios registrados más activos es patrullar y revertir acciones vandálicas. La utilización de bots que marcan posibles ediciones vandálicas agiliza el proceso de reversión, no sólo por la rapidez, sino también porque hace que usuarios con conocimientos medios puedan realizar reversiones que de otro modo deberían ser realizadas por expertos del dominio. Es de señalar que el porcentaje de artículos que nunca ha sido revertido es muy elevado (67%), y otro 13% lo ha sido solamente una vez, lo cual indica que las principales acciones de vandalismo recaen sobre un número no muy elevado de artículos.

Para un porcentaje alto de usuarios se ha obtenido el país de origen, lo cual ha permitido contabilizar la contribución que se realiza desde cada uno de ellos. Este aporte es importante, pues son pocos los estudios específicos de sistemas de información en línea que obtienen información del origen de los contribuyentes, no ya por la dirección IP (por otro lado bastante frecuentes en analizadores de visitas web), sino por el sistema de representación que tiene Wikipedia en lo referente a las categorías que pueden asignarse los propios usuarios para indicar su país o región de origen.

6. Agradecimientos

Este trabajo ha sido parcialmente financiado por la Fundación Memoria de D. Manuel Solórzano Barruso de la Universidad de Salamanca, proyecto de Ref. FS/5-2014.

Referencias

Adler, B.T.; De Alfaro, L.; Kulshreshtha, A.; Pye, I.; (2011a). Reputation systems for open collaboration. // Communications of the ACM, 54:8, 81-87.

Adler, B.T.; De Alfaro, L.; Mola-Velasco, S.M.; Rosso, P.; West, A.G. (2011b). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features, LNCS: Computational Linguistics and Intelligent Text Processing, 6609, 277-288.

Baeza-Yates, R.; Castillo, C., Lopez, V. (2005): Characteristics of the Web of Spain. Cybermetrics, 9(1).

Baeza-Yates, R.; Castillo, C., y Graells, E. (2006): Características de la Web Chilena. Universidad de Chile.

Berrocal, J. L. A.; Figuerola, C. G.; Zazo, Á. F. (2004). Cibermetría: nuevas técnicas de estudio aplicables al Web. Ed. Trea.

Fernández-Molina, J. C.; Guimarães, J. A. C. (2002). Ethical aspects of knowledge organization and representation in the digital environment: their articulation in professional codes of ethics. A// dvances in Knowledge Organization, 8, 487-492.

Francke, H.; Sundin, O. (2010). An inside view: credibility in Wikipedia from the perspective of editors. // Information Research, Special Supplement: Proceedings of the 7th International Conference on Conceptions of Library and Information Science, London. 15:3.

Gabrilovich, E.; Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. // Journal of Artificial Intelligence Research. 34, 443-498.

Gandica, Y.; Carvalho, J.; dos Aidos, F. S. (2015). Wikipedia editing dynamics. // Physical Review E. 91:1, 012824.

Geiger, R. S.; Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. // Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work. 117-126.

Hu, M.; Lim, E. P.; Sun, A.; Lauw, H. W.; Vuong, B. Q. (2007). Measuring article quality in Wikipedia: models and evaluation. // Proceedings of the sixteenth ACM Conference on Information and Knowledge Management. 243-252.

Kimmons, R. M. (2011). Understanding collaboration in Wikipedia. First Monday. 16:12.

Leskovec, J.; Huttenlocher, D.; Kleinberg, J. (2010) Governance in social media: A case study of the Wikipedia promotion process. // Proceedings of the International Conference on Weblogs and Social Media, ICWSM'10.

Lewandowski, D.; Spree, U. (2011). Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? // Journal of the American Society for Information Science and Technology. 62:1, 117-132.

Luyt, B.; Aaron, T. C. H.; Thian, L. H.; Hong, C. K. (2008): Improving Wikipedia's accuracy: Is edit age a solution? //

Journal of the American Society for Information Science and Technology. 59:2, 318–330.

Magnus, P. D. (2009). On trusting Wikipedia. // Episteme: A Journal of Social Epistemology. 6:1, 74-90.

Medelyan, O.; Milne, D.; Legg, C.; Witten, I. H. (2009). Mining meaning from Wikipedia. // International Journal of Human-Computer Studies. 67:9, 716-754.

Mesgari, M.; Okoli, C.; Mehdi, M.; Nielsen, F. Å.; Lanamäki, A. (2015). The sum of all human knowledge: A systematic review of scholarly research on the content of Wikipedia. // Journal of the Association for Information Science and Technology. 66, 219–245.

Müller, C.; Gurevych, I. (2009). Using wikipedia and wiktionary in domain-specific information retrieval. // Evaluating Systems for Multilingual and Multimodal Information Access. Berlin, Heidelberg: Springer. 219-226.

Okoli, C.; Mehdi, M.; Mesgari, M.; Nielsen, F. Å.; Lanamäki, A. (2012). The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. // Social Science Research Network, 2021326.

Pehcevski, J.; Thom, J. A.; Vercoustre; A. M.; Naumovski, V. (2010). Entity ranking in Wikipedia: utilising categories, links and topic difficulty prediction. // Information Retrieval, 13:5, 568-600.

Priedhorsky, R.; Chen, J.; Lam, S. T. K.; Panciera, K.; Terveen, L.; Redl, J. (2007). Creating, destroying, and restoring value in Wikipedia. // Proceedings of the 2007 international ACM Conference on Supporting Group Work (pp. 259-268). ACM.

Sepehri, H.; Makazhanov, A.; Rafiei, D.; Barbosa, D. (2012). Leveraging editor collaboration patterns in Wikipedia. // Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12. 13-22.

Soliman, M.; Gourdain, P. (2008). La revolución Wikipedia. Madrid: Alianza Editorial, ISBN: 978-84-206-8236-5.

Stuart, D. (2013). Web metrics for library and information professionals. London: Facet Publ.

Wu, G., Harrigan, M., Cunningham, P. (2011) Characterizing Wikipedia pages using edit network motif profiles. // Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents.45-52.

Wu, Q.; Irani, D.; Pu, C.; Ramaswamy, L. (2010). Elusive vandalism detection in Wikipedia: a text stability-based approach. // Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10. 1797-1800.

Yasseri, T.; Kertész, J. (2013). Value production in a collaborative environment. // Journal of Statistical Physics. 151:3-4, 414-439.

Enviado: 2015-04-12. Segunda versión: 2015-08-10.
Aceptado: 2015-08-17.