

EL PROYECTO DE COLABORACIÓN DE LA BIBLIOTECA DE LA UNIVERSIDAD COMPLUTENSE DE MADRID CON GOOGLE BÚSQUDA DE LIBROS

Palafox, Manuela

Acebes, Ricardo

Biblioteca de la Universidad Complutense de Madrid

mpalafox@buc.ucm.es / racebes@buc.ucm.es

Resumen:

La firma del acuerdo entre la Universidad Complutense y Google en 2006 ha permitido iniciar un proyecto de digitalización masiva de las colecciones libres de derechos de autor que custodia la Biblioteca Complutense. Desde 2008, miles de obras de dominio público pueden consultarse libre y gratuitamente en Internet, desde Google Búsqueda de Libros y desde 2009 desde el Catálogo de la Biblioteca. El documento describe las fases del proyecto de digitalización y enumera las actividades que se han desarrollado durante este proceso.

Palabras clave: Proyectos de digitalización masiva; Bibliotecas universitarias; Google Búsqueda de Libros.

Abstract:

The agreement between Complutense University and Google has allowed a large-scale digitization project of the complutense collections out of copyright. From 2008, thousands of books in the public domain have free access with an Internet connection and from 2009 it is possible to access from the Complutense Library Catalogue. The document describes the digitization project steps and the activities carried out during the process.

Keywords. Massive digitization projects; Academic Universities; Google Book Search

Introducción

Como se sabe, las bibliotecas que tienen una cierta antigüedad no sólo se dedican a la gestión de la información y la documentación actual y a tratar de ofrecer servicios eficientes para satisfacer este tipo de demanda, sino que han de prestar atención igualmente a las colecciones que han ido acumulando y organizando desde sus orígenes y que pueden estar compuestas por manuscritos, libros y revistas de gran valor en diferentes soportes, como el papel y el pergamino.

Como es lógico, un porcentaje elevado de estas colecciones que se van alejando en el tiempo retroceden puestos en el interés primario de los lectores. Pero éstos también se encuentran con las dificultades que, para su consulta, añaden las bibliotecas, por motivos de conservación y, en algunos casos, de organización.

A decir verdad, hasta ahora no existía un sistema realmente eficaz para conjugar la preservación, la conservación y la difusión de estas colecciones. Pero esta situación ha cambiado gracias a la aparición y extensión de la era de los ordenadores, el formato digital e Internet. Y ha ocurrido, no por la mera existencia de estas opciones tecnológicas, sino porque esta tecnología se encuentra integrada, y cada vez lo irá estando más, con un conjunto amplio de mecanismos de socialización, con las formas que utilizamos para relacionarnos y comunicarnos. Si alguien en cualquier parte del mundo dispone de acceso a un ordenador conectado a Internet o de un teléfono móvil, porque los utiliza para trabajar, para estudiar, para informarse o para comunicarse, entonces también tendrá la oportunidad de leer un libro del siglo XVI, siempre que éste se encuentre digitalizado y disponible en la red.

La Universidad Complutense de Madrid, convencida de las posibilidades que se abren por esta vía, lleva tiempo en el empeño de organizar y difundir a través de Internet tanto la producción científica y académica de su personal docente e investigador, como de ofrecer al público en general la ocasión de estudiar el contenido de su patrimonio documental y cultural.

Entre las diferentes acciones llevadas a cabo en esta línea se encuentran nuestras propias experiencias previas en digitalización y la puesta en funcionamiento de un repositorio institucional de e-prints¹ y de un portal con las revistas científicas publicadas por la Universidad². Además, ésta se ha comprometido con los principios y actuaciones promovidos por el movimiento internacional a favor del acceso abierto, adhiriéndose en el verano de 2006 a la *Declaración de Berlín*. Y también en ese año firmó un acuerdo de colaboración con la empresa Google, para la digitalización de los fondos bibliográficos de la biblioteca que son de dominio público. Se trata de un acuerdo ambicioso y del que no existían precedentes en España y sólo unos pocos en el mundo.

Otras experiencias en digitalización y acceso abierto

A mediados de la década de 1990, aprovechando el desarrollo de las nuevas tecnologías, la Biblioteca Complutense inició un proyecto pionero de digitalización de fondos bibliográficos antiguos en colaboración con la Fundación Ciencias de la Salud y los Laboratorios GlaxoSmithKline. Se trata del *Proyecto Dioscórides*, cuyo principal objetivo era ofrecer acceso público al fondo histórico de biomedicina de la biblioteca. Posteriormente, se amplió el ámbito de digitalización a otras materias, formándose así la *Biblioteca Digital Dioscórides*, que actualmente contiene cerca de 3.000 libros del siglo XV al XIX, manuscritos e impresos, y aproximadamente 50.000 grabados en acceso libre en Internet. Es un proyecto abierto y se siguen digitalizando libros a día de hoy.

No obstante, los resultados obtenidos hasta el año 2006, muy limitados pese a los esfuerzos, nos mostraron algunas conclusiones: por un lado, el ritmo de digitalización era muy bajo, por lo que no era posible extender el proyecto a todo nuestro patrimonio bibliográfico histórico, ya que no dispondríamos de las obras digitalizadas en un plazo de tiempo razonable. Por otro, las restricciones presupuestarias de la biblioteca no nos permitían invertir en los desarrollos tecnológicos y en los recursos humanos y de otros tipos que necesitábamos. En consecuencia, era preciso buscar la manera de continuar la labor ya realizada con un plan de digitalización masiva a través de una iniciativa comercial, como la que estaba acometiendo Google con su programa Búsqueda de Libros.

Recientemente, la biblioteca ha obtenido fondos del Ministerio de Cultura para la digitalización de algunas obras que por sus características no pueden ser digitalizadas por Google. También se ha firmado un convenio de colaboración con la empresa Extramuros Edición para digitalizar y realizar la edición facsimilar de algunas obras libres de derechos de autor de nuestra colección. Al igual que en el caso anterior, este proyecto nos permitirá digitalizar obras que, por el momento, no pueden ser digitalizadas por Google. También hay un proyecto de difusión de los fondos antiguos y otras colecciones singulares de la Biblioteca Complutense en el Portal de la Biblioteca Virtual Miguel de Cervantes.

El otro frente de actuación en relación con el acceso abierto en nuestra biblioteca no se orientaba ya tanto al patrimonio bibliográfico que conserva, sino a la producción científica y académica de la Universidad Complutense. En primer lugar, se inició la digitalización de las tesis doctorales y las revistas científicas editadas por la Universidad. A continuación, se desarrollaron dos aplicaciones para la difusión en Internet de estos dos grupos de contenidos: el Archivo Institucional E-Prints Complutense³ y el Portal de Revistas Científicas Complutenses⁴. Con ambos proyectos, la biblioteca se proponía incrementar la visibilidad, el uso y el impacto de la investigación de la Universidad, así como mejorar su organización, acceso y preservación.

¹ <http://eprints.ucm.es>

² <http://www.ucm.es/BUCM/revistasBUC/portal/modulos.php?name=principal&col=1>

³ E-Prints Complutense: <http://eprints.ucm.es>

⁴ Portal de Revistas Científicas Complutenses:
<http://www.ucm.es/BUCM/revistasBUC/portal/modulos.php?name=principal&col=1>

Los beneficios del acuerdo entre la Universidad Complutense de Madrid y Google

Como se ha dicho anteriormente, en septiembre de 2006 la Universidad Complutense de Madrid y Google firmaron un acuerdo de cooperación centrado básicamente en la digitalización masiva de las colecciones de la biblioteca libres de derechos de autor, que podrán consultarse libre y gratuitamente en Internet, al mismo tiempo que se mejora la conservación y preservación de los materiales bibliográficos originales. Junto a estos dos factores, el acceso y la preservación, la Biblioteca Complutense tiene previsto usar los materiales digitales de sus fondos para la investigación creando nuevos servicios para nuestros investigadores, especialmente de Humanidades. Utilizando los textos OCR se podrá comparar un texto seleccionado o analizar las repeticiones de frases o palabras entre distintos textos. Para ello, será fundamental el trabajo de cooperación entre los investigadores, informáticos y bibliotecarios complutenses. Algunas bibliotecas socias del proyecto Google Búsqueda de Libros, como Harvard o Stanford, están explorando estas vías. Como señala Flecker, de la Universidad de Harvard, “la posibilidad de que la colección de libros digitalizados esté disponible en el futuro para llevar a cabo minería de textos descubrirá nuevos caminos de hacer investigación⁵”.

En el acuerdo con Google, la biblioteca aporta sus fondos y el personal técnico que realiza las tareas necesarias para la selección y preparación de las obras para digitalizar, así como los trabajos de supervisión de la integridad de las mismas, mientras que Google se hace cargo de todos los costes de la digitalización, de la instalación del espacio dedicado a este proceso y del transporte de los materiales.

Existen dos copias de los materiales digitalizados, una que conserva la Universidad y otra de Google, que se podrán consultar y manejar libremente, respetando los derechos de autor, desde “Google Búsqueda de Libros” y desde la Web y el catálogo de la Biblioteca Complutense. No obstante, la biblioteca podrá disponer de su copia para otros objetivos, según le convenga para sus intereses, como la cooperación en proyectos bibliotecarios con otras instituciones, utilizando solamente un porcentaje de las imágenes y exceptuando actividades con ánimo de lucro.

Los fondos de la Universidad Complutense se suman así a un proyecto internacional amplio, “Google Búsqueda de Libros”, que permite el acceso a millones de libros libres de derechos de autor, a través de una tecnología puntera que facilita la recuperación mediante búsquedas en el texto completo del contenido y en múltiples índices y metadatos. De este modo, se estima que el acceso a las colecciones de nuestra Universidad, junto con las de algunas de las principales bibliotecas del mundo, mejorará significativamente, al tiempo que, como resultado de su presencia en Internet, se da un paso muy importante en el proceso de difusión abierta y sin restricciones del saber y en el de la preservación de la memoria cultural de la humanidad depositada en los libros.

De esta forma, por mediación de este acuerdo, nuestra biblioteca mejora considerablemente sus herramientas para el desarrollo cultural, científico e investigador. También se sitúa en una posición estratégica de colaboración con algunas de las principales bibliotecas del mundo y, además, se alía con la empresa líder del sector de la información, más allá incluso de la mera actualización y mejora del proyecto de digitalización que había iniciado hacia algunos años.

Adicionalmente, el proyecto de digitalización tiene también otros beneficios paralelos, relacionados con la información sobre las colecciones que obtenemos en el proceso y las posibilidades de aceleración de algunos trabajos bibliotecarios, como la finalización de la catalogación de los fondos históricos.

Por la parte de Google, el beneficio más visible es la incorporación a su proyecto de búsqueda de libros en Internet de las colecciones de nuestra biblioteca. Como es bien conocido, se trata de un proyecto constituido por dos programas: el Programa de Socios y el Programa de Bibliotecas.

A través del *Programa de Editoriales*, Google firma acuerdos con editoriales y autores, que le aportan los libros para digitalizarlos y hacerlos accesibles en línea, si bien en un número limitado de páginas o fragmentos relevantes, para que cualquier persona pueda, si son de su interés y así lo desea, comprarlos u obtenerlos en una biblioteca cercana. Google digitaliza los libros de forma gratuita y no obtiene ningún beneficio por la compra de libros.

⁵ Grogg, J.E. y Ashmore, B. (2007)

Con el *Programa de Bibliotecas* pretende, por su parte, digitalizar millones de libros almacenados en las principales bibliotecas de EE.UU, Europa y Asia, cuyo número se acerca actualmente a treinta. Las últimas bibliotecas que se han incorporado al proyecto son la Biblioteca de la Universidad de Keio, Japón y la Biblioteca Municipal de Lyon, Francia. La intención de Google es trabajar con cada una de estas bibliotecas y desarrollar un flujo de trabajo de digitalización masiva a una escala sin precedentes hasta hoy, para organizar la información mundial y hacer que sea universalmente accesible de forma gratuita, facilitando así, además, la búsqueda y la lectura de libros relevantes difíciles de conseguir de otro modo, por su carácter único, por encontrarse a miles de kilómetros de los posibles lectores o porque se hallan descatalogados.

Los objetivos del proyecto de digitalización masiva

Teniendo esto en cuenta, los objetivos concretos que se ha marcado la Biblioteca de la Universidad Complutense de Madrid al iniciar esta colaboración con Google se pueden resumir en los siguientes:

- Ofrecer acceso al patrimonio bibliográfico de la Universidad y contribuir a su difusión, acceso y preservación.
- Aumentar el uso de la colección, ya que la experiencia confirma que incluso las versiones en papel de los libros digitalizados son más usadas que las de aquellos, de épocas y temáticas similares, que no han sido escaneadas.
- Ofrecer un sistema de información de calidad y conseguir que el conocimiento de la Universidad revierta en la sociedad.
- Catalogar los fondos antiguos y del siglo XIX pendientes de catalogar.
- Establecer un plan de conservación y restauración de libros dañados.
- Utilizar los materiales digitales en proyectos de investigación y desarrollo.

Diseño y planificación del proyecto de digitalización

Tras la firma del acuerdo hubo de constituirse un grupo de trabajo integrado por personal bibliotecario e informático de la Universidad, tanto de los Servicios Centrales de la Biblioteca como de las bibliotecas de las Facultades y Escuelas de nuestra universidad. Sus funciones fueron el establecer las líneas directrices del proyecto, dirigirlo, controlar su progreso y elaborar los documentos necesarios que informaran sobre su desarrollo.

Durante la fase del diseño del proyecto se realizaron las actividades que se presentan a continuación:

Análisis de la colección

En primer lugar, se inició un análisis de la colección con el objeto de conocer el número de ejemplares libres de derechos de autor, según la legislación española, que posee la biblioteca y que, por lo tanto, eran susceptibles de digitalizar.

Debe tenerse en cuenta que la Biblioteca de la Universidad Complutense de Madrid cuenta con más de 30 bibliotecas especializadas, que dan servicio a sus respectivas Facultades y Escuelas Universitarias, incluyendo una biblioteca de fondo antiguo o histórico. Esta biblioteca de fondo histórico administra y conserva la mayor parte de las colecciones de materiales antiguos, hasta los inicios del siglo XIX, mientras que las obras restantes del siglo XIX se encuentran repartidas en las demás bibliotecas.

Se realizaron recuentos de ejemplares por bibliotecas y se obtuvieron distintas listas, clasificadas por tipos de documentos y ubicaciones. Se cuantificaron o se estimaron los libros que aún no estaban catalogados en cada una y se observó que las bibliotecas con mayor volumen de libros de dominio público son la Biblioteca Histórica y, a continuación, las de las Facultades de Filología, Derecho, Medicina, Farmacia, Veterinaria, Geografía e Historia, Educación, etc.

Informes de situación

Tras el análisis de la colección, se hicieron diversas inspecciones de estas bibliotecas, con la intención de preparar un informe que recogiera datos sobre las instalaciones de los depósitos donde se encuentran los fondos antiguos y sus características de accesibilidad y posibles dificultades puntuales. También se describían brevemente las colecciones, su sistema de ordenación y su estado de conservación general, y se estimaba el número de obras pendientes de catalogar. Por último, se especificaban las tareas necesarias previas y posteriores a la selección y el personal necesario en cada caso.



Depósito de la Biblioteca de Medicina



Depósito de la Biblioteca de Farmacia



Depósito de la Biblioteca de Derecho

Plan de catalogación

En diciembre de 2006, varios servicios centrales de la biblioteca elaboraron un plan de catalogación para los fondos del siglo XIX pendientes de procesar.

Asimismo, se creó un grupo de diez catalogadores, al que se dio una formación específica, y otro de personal de apoyo para los trabajos de movimientos de fondos, comprobación de la existencia o no de registros bibliográficos en el catálogo y, en su caso, en otros catálogos, vía Z39.50, para la creación de los registros necesarios.

Por otra parte, en cada biblioteca con colecciones del siglo XIX, el personal responsable del proceso técnico seleccionó los materiales que había que catalogar y una persona se encargó de supervisar todo el trabajo, siguiendo para ello un procedimiento creado al efecto. Además, el servicio central de proceso técnico y normalización realizó el oportuno control de autoridades.

Durante el año 2007 se catalogaron todos los fondos correspondientes al siglo XIX y en 2008 se ha continuado la catalogación intensiva de los fondos de cualquier época pendientes de catalogar de la Biblioteca Histórica. A fecha de junio de 2008 se han catalogado más de 70.000 ejemplares. En 2009 se finalizará la catalogación del resto del fondo histórico de la biblioteca.

Guía de criterios de selección

Igualmente, ha sido necesario definir unas pautas para la selección de los materiales para digitalizar. La guía elaborada para ello, establece una serie de puntos que se revisan con el ejemplar en mano y que permiten evaluar, con criterios objetivos y cuantificables, las condiciones de conservación, entre otras.

Programa de digitalización

A mediados de 2007 se analizaron y describieron por escrito los flujos de trabajo y las operaciones de logística, esto es, básicamente de movimiento de libros que debían seguirse cuando el proyecto de digitalización de los fondos entrara en la fase de producción. A continuación, se diseñó el programa de digitalización, con la secuencia y un cronograma detallado de los diferentes procesos por bibliotecas.

Desarrollos tecnológicos

Dado que se trata de un proyecto determinado en gran medida por la tecnología informática, la biblioteca estudió detenidamente qué necesidades se iban a presentar que debiéramos y pudiéramos satisfacer por nuestros propios medios, unidas a los desarrollos que nos ofrecía Google. De esta forma, podemos señalar dos fases:

Fase de digitalización

- El personal informático vinculado al proyecto ha desarrollado dos aplicaciones para realizar la gestión interna del proceso de digitalización. En primer lugar, una aplicación Web que permite disponer en línea y en tiempo real de la información actualizada y precisa de todos los procesos implicados en la digitalización, con datos concretos y estadísticas. En esta aplicación se encuentran almacenados los metadatos de los libros incluidos en el programa de digitalización, procedentes del catálogo de la biblioteca. También ofrece una visión completa de los movimientos diarios de los libros en las bibliotecas y del envío a Google de los libros seleccionados.
- Una segunda aplicación, conectada a la anterior, se utiliza directamente en los depósitos a través de una PDA. Para empezar, se lee el código de barras de un libro y, a continuación, se presenta en la pantalla táctil un formulario donde el bibliotecario que realiza la selección señala una serie de características de los libros respecto a su estado de conservación. Después, esta información se transfiere a la aplicación Web y, por último, al catálogo de la biblioteca, donde se incluye en una nota interna los siguientes datos:
 - Los libros que sobrepasan el tamaño para poder ser digitalizables, y se anotan las dimensiones (alto, ancho y grosor).
 - También se registran datos relativos al estado del bloque de hojas (presencia de hongos, grado de deterioro físico, hojas sueltas, papel frágil por acidez e intonso) y a la encuadernación del libro (valiosa, débil, perdida, necesidad de reencuadernación, no apertura del libro y encuadernación deteriorada).



- Gracias a este proceso, la biblioteca dispondrá, al final del proyecto de digitalización masiva, de una información muy exhaustiva y valiosa sobre el estado de conservación de su colección histórica.
- Por último, los servicios informáticos también tienen prevista la adquisición de un servidor con la memoria necesaria para almacenar todas las imágenes resultantes del proceso de digitalización.

Fase posterior a la digitalización

Una vez digitalizados los libros, se puede acceder a ellos a través de: Google Búsqueda de Libros (Google Book Search), de una interfaz de Google específica para la búsqueda de libros complutenses⁶ y a través de los registros bibliográficos del catálogo de la biblioteca.

En una fase posterior, se podrá desarrollar una interfaz de acceso y un sistema de recuperación que permita la consulta de la copia digital de los libros alojados en los servidores de la Universidad Complutense.



Preservación de los contenidos digitales

Según la Política de Gestión de las Colecciones de la Biblioteca de la Universidad Complutense⁷, “la Biblioteca Complutense mantendrá una política de preservación de recursos digitales, constituida por el conjunto de actividades e intervenciones requeridas para garantizar, a largo plazo, la accesibilidad y legibilidad de los objetos o documentos digitales fidedignos que la Universidad requiere para sus fines de aprendizaje, docencia, investigación y demás actividades relacionadas con sus objetivos institucionales”. En el apartado Ámbito de la preservación digital se incluyen, entre otros, los documentos en soporte digital correspondientes al Patrimonio Bibliográfico de la UCM.

Unas de las motivaciones principales de la UCM a la hora de la firma del acuerdo con Google en 2006 era el deseo de asegurar que los materiales de la Biblioteca permanecieran accesibles para las generaciones futuras. Si los usuarios acceden a las copias digitales, en lugar de a los materiales analógicos, se minimiza el manejo de los libros originales y, de esta forma, se está realizando una función de preservación. También cumple una función de preservación si se utilizan las copias digitales como *backups* y se establecen directrices según los estándares y especificaciones técnicas necesarias para asegurar el acceso en el futuro.

En cuanto a las especificaciones de digitalización, el formato de conservación de las imágenes de la copia digital de la Biblioteca es JPEG2000. El formato JPEG2000 se está utilizando en muchos proyectos de digitalización masiva debido a su técnica de compresión que produce ficheros de tamaño menor que otros y que, por tanto, son más eficientes para almacenarlos, procesarlos y transferirlos⁸. La Biblioteca Complutense está utilizando en todos sus proyectos de digitalización los metadatos de preservación PREMIS⁹ estructurados según el esquema METS (Metadata Encoding & Transmisión), siguiendo la recomendación del Ministerio de Cultura de España¹⁰.

⁶ <http://www.ucm.es/BUUCM/atencion/25403.php>

⁷ Documento en proceso de aprobación.

⁸ Rieger, Oya Y. (2008)

⁹ <http://www.loc.gov/standards/premis/>

¹⁰ http://travesia.mcu.es/documentos/pautas_digitalizacion.pdf

Conclusiones

Gracias a este proyecto, la Biblioteca de la Universidad Complutense de Madrid está habilitando el acceso libre a un gran número de libros, reduciendo drásticamente la dificultad de la distancia y de la presencia, condiciones ineludibles de las obras impresas y, por lo tanto, ampliando el número de potenciales lectores; pero también reduciendo los riesgos para la integridad de las obras antiguas.

Por otra parte, este acuerdo ya está beneficiando a la biblioteca universitaria de un modo único, favoreciendo el desarrollo paralelo de otros proyectos interesantes, como la revisión del estado de las colecciones, tanto desde el punto de vista físico, como desde el relativo al trabajo bibliotecario, la catalogación de fondos antiguos y del siglo XIX aún no procesados, la encuadernación de obras que carecían de cubiertas, etc.

Además, el proyecto de digitalización ha permitido al personal de la Biblioteca Complutense trabajar en un proyecto cooperativo con importantes bibliotecas universitarias de todo el mundo y con Google, empresa líder en el sector de la información.

Referencias

Bjørner, S. (2007) "Complutense University of Madrid: Different Language, Similar experience". *Searcher* (April). Disponible en: http://www.infotoday.com/searcher/apr07/Grogg_Ashmore.shtml. [Consulta: 10/02/09]

Grogg, J.E. y Ashmore, B. (2007) "Google Book Search Libraries and Their Digital Copies". *Searcher* (April). En: http://www.infotoday.com/searcher/apr07/Grogg_Ashmore.shtml [Consulta: 10/02/09]

Ministerio de Cultura. Subdirección General de Coordinación Bibliotecario. (2008). Directrices para proyectos de digitalización de colecciones y fondos de dominio público, en particular para aquellos custodiados en bibliotecas y archivos. Apéndice A. En: http://travesia.mcu.es/documentos/pautas_digitalizacion.pdf. [Consulta: 10/02/09]

Rieger, Oya Y. (2008). "Preservation in the Age of Large-Scale Digitization. A White Paper". Washington. Council on Library and Information Resources.

Toobin, J. (2007). "Google's Mopon Shot: The Quest for the Universal Library". *The New Yorker* (February, 5). En: http://www.newyorker.com/reporting/2007/02/05/070205fa_fact_toobin. [Consulta: 10/02/09]