

CAPTURAR LA WEB VASCA PARA OFRECER CONTENIDOS DIGITALES EN LÍNEA

Francisca Pulgar Vernalte(*), responsable del Servicio de Bibliotecas, Gobierno Vasco, f-pulgar@ej-gv.es ; Sonia Marcos Maciá, Técnico Documentalista. smmacia@gmail.com

Resumen: Los cambios incesantes en los soportes para la transmisión del conocimiento hacen imprescindible que las bibliotecas deban adaptarse obligatoriamente a las tendencias que imponen tanto los usuarios como la aparición de nuevas herramientas en Internet. Elementos como la web 2.0 con toda una serie de productos innovadores y fáciles para los usuarios, así como las aplicaciones de software libre han venido a aumentar las posibilidades de los bibliotecarios para acercar la información a la ciudadanía. En esta línea, desde el Departamento de Cultura del Gobierno Vasco se inició en el 2007 el desarrollo de un interesante proyecto: el archivo electrónico del patrimonio digital vasco, conocido como Ondarenet (www.euskadi.net/ondarenet). Este archivo tiene la función de capturar, conservar y difundir el patrimonio que nace ya en formato digital y que trata de temas relacionados con la cultura, sociedad, la lengua o el País Vasco.

Palabras clave: archivo web, patrimonio digital, captura web, Euskadi

1. Introducción

Elementos como la web 2.0 con toda una serie de productos innovadores y fáciles de utilizar por parte de los usuarios, así como las aplicaciones de software libre están cambiando los modos de interrelación entre bibliotecas y usuarios, pero además ofrecen a los bibliotecarios nuevos escenarios para acercar la información a la ciudadanía.

Desde los foros profesionales se insiste en la necesidad de que las bibliotecas se sitúen y posicionen en la red, que creen contenidos, y que busquen al usuario en el medio en el que este se desenvuelve.

Pero en esta interrelación biblioteca ciudadanía no se debe olvidar que la popularización de Internet, la facilidad de las herramientas que nos ofrece para la edición y la rapidez y alcance de difusión, han contribuido a un incremento en la publicación y edición de contenidos en la red, alejándonos cada vez más de los formatos tradicionales.

En esta línea en un artículo reciente, titulado “La edición científica tradicional frena la difusión del saber” se reconocía que la publicación en revistas de acceso abierto aumenta el impacto de los artículos publicados frente a la edición en papel (1).

Los cambios incesantes referidos a la edición de los soportes para la transmisión del conocimiento hacen imprescindible que las bibliotecas deban adaptarse obligatoriamente a las tendencias que imponen tanto los usuarios como la aparición de nuevas herramientas en Internet. De ahí que cada biblioteca o institución, según su misión y tipología de usuarios, esté obligada a

definir su propia estrategia sobre qué tipo de contenido digital debe elaborar y publicar en Internet.

Esta comunicación se elabora desde la perspectiva de lo que será la futura biblioteca de Euskadi, por lo que el enfoque de la misma serán acordes a las funciones atribuidas a una biblioteca nacional en referencia al patrimonio digital, y más exactamente a los recursos que nacen ya en formato digital, como es el caso de las páginas web.

2. Capturando la web

Ya en la introducción se comentaba que la realidad de la publicación está cambiando. Es un hecho contrastable que la edición en el tradicional soporte en papel se ha estancado –e incluso descendido- frente al espectacular aumento que se produce en el registro de dominios, o lo que es lo mismo, en formatos digitales solo accesibles a través de internet.

Las cifras extraídas en dos fuentes representativas, como son Red.es (2) y el Ministerio de Cultura (3) no dejan lugar a dudas sobre esta evolución. En el cuadro adjunto se puede comprobar cómo entre 2003 y 2008 la producción de libros en soporte papel se ha mantenido constante con cierta tendencia a la baja, mientras que el registro de dominios .es ha ido creciendo, especialmente desde 2005, año en el que se contempla un aumento espectacular en el número de dominios .es registrados.

CONCEPTO	AÑOS					
	2003	2004	2005	2006	2007	2008
Producción de libros en soporte papel	77.950	77.367	76.265	77.330	75.006	73.275
Registro de dominios. es	27.682	14.151	213.291	209.274	297.453	277.430

Tabla 1. Cuadro comparativo entre publicación en soporte papel y registro de dominios en España (2003-2008)

Según las últimas proyecciones realizadas, el universo digital se ampliará diez veces de los 116 exabytes* en 2006 a 1.800 exabytes en 2011. Lo que significa que la producción actual de contenidos digitales, supera con creces nuestra capacidad para mantenerla. La tendencia refleja que para 2011, la mitad de contenido creado en línea será inaccesible, habrá desaparecido (4).

En este sentido mencionaremos la aparición de un nuevo elemento: el “digital dark age”, que aborda el problema de la recopilación y conservación de la información almacenada en archivos digitales, una información que, a menudo, se pierde por la rápida desaparición de los recursos o por la obsolescencia de unos sistemas que cambian rápidamente (5).

* 1 Exabyte = mil millones de Gigabytes

XV JORNADAS BIBLIOTECARIAS DE ANDALUCÍA

A medida que nuestra “memoria colectiva” se encuentre cada vez con mayor frecuencia en línea, crece el peligro de que contenidos, acontecimientos y contextos relevantes que han ocurrido hace, relativamente poco tiempo, se pierdan de manera irreversible.

Es necesario que instituciones con experiencia y recursos tomen cartas en el asunto y elaboren políticas y estrategias encaminadas a conservar esta parte de la memoria colectiva. En la “Carta para la preservación del patrimonio digital de la UNESCO” se habla de la “necesidad de pasar a la acción”

“A menos que se haga frente a los peligros actuales, el patrimonio digital desaparecerá rápida e ineluctablemente. El hecho de estimular la adopción de medidas jurídicas, económicas y técnicas para salvaguardar ese patrimonio redundará en beneficio de los propios Estados Miembros. Urge emprender actividades de divulgación y promoción, alertar a los responsables de formular políticas y sensibilizar al gran público tanto sobre el potencial de los productos digitales como sobre los problemas prácticos que plantea su preservación (6)

Es aquí donde entran en juego archivos y bibliotecas, acostumbradas a gestionar grandes volúmenes de información y preparadas para los cambios tecnológicos. No es casual que los grandes proyectos relacionados con la preservación de sitios web estén liderados por bibliotecas nacionales, es el caso de proyectos como Kulturarw³ (Biblioteca Nacional de Suecia), Pandora (Biblioteca Nacional de Australia) o Minerva (Library of Congress).

Pero para la buena marcha de este tipo de proyectos es necesaria la cooperación interinstitucional. Por un lado, universidades e instituciones relacionadas con la investigación están demostrando gran interés por la recolección y acceso a contenidos digitales a través de los Repositorios OAI. Pero, además, se hace imprescindible llegar a acuerdos con las entidades o individuos creadores de los contenidos con el fin de evitar problemas legales relacionados con el copyright y el depósito legal.

3. Ondarenet: el archivo electrónico de la web vasca

El panorama descrito hasta ahora hace evidente que las bibliotecas nacionales tienen por delante un cometido mucho más amplio que el de recoger sólo la producción bibliográfica de un país, ya que si quieren conformar una auténtica colección nacional tienen que abordar irremediabilmente la recopilación y preservación de los recursos digitales surgidos en la red.

De ahí que, tal y como hemos comentado, sean las propias bibliotecas nacionales las que lideran los principales proyectos relacionados con la conservación, preservación y difusión del patrimonio digital, y en nuestro caso, correspondería a la Biblioteca de Euskadi el desarrollo de este tipo de acciones.

Si bien es cierto, que aún no existe un edificio que albergue dicha institución, hay que reconocer que la Ley 11/2007, de 26 de octubre, de bibliotecas de Euskadi, supone un avance al respecto, ya que en su artículo 27 crea la Biblioteca de Euskadi, y además en el mismo artículo, apartado 4, especifica

que “la Biblioteca de Euskadi se constituye en sede del patrimonio digital vasco”.

La importancia de esta ley se refleja tanto a la hora de establecer los principios legales que permitan recoger y difundir el patrimonio digital vasco, como al hacer referencia a la definición de obra bibliográfica, en la que se incluyen los formatos digitales. En el artículo 33 del título VI, referido al depósito bibliográfico de Euskadi, define la obra bibliográfica como “toda obra presentada para su uso o difusión, tanto en formato analógico como digital y en soporte papel, electrónico o de cualquier otro tipo que pueda crearse en el futuro (...) (7)

Además, hasta que se concrete la ubicación de esta importante infraestructura cultural, la misma Ley adscribe las tareas y funciones de dicha biblioteca al Departamento de Cultura del Gobierno Vasco, y concretamente, al Servicio de Bibliotecas.

En este sentido, desde el Departamento de Cultura se han puesto en marcha una serie de iniciativas encaminadas a la recopilación y preservación del patrimonio digital vasco, iniciativas que se recogieron en el Plan Director de digitalización, preservación y difusión del Patrimonio Cultural Vasco⁸, elaborado en el 2005. Entre las líneas de actuación previstas en dicho Plan se incluía la “Definición de los modelos y funciones a desarrollar por las Administraciones Públicas en el ámbito de la preservación y acceso al patrimonio digital”, resumidas en tres puntos fundamentales:

1. *Implantación del modelo óptimo de almacenamiento y preservación del patrimonio digital.* Orientado hacia la preservación a largo plazo de los activos digitales, en el que se decidirá sobre modelo tecnológico, estándares, soportes físicos, normativa a aplicar, modelos de negocio, viabilidad, etc.
2. *Puesta en marcha de la política de preservación web.* Se trata de concretar e implantar la política de preservación web, como aspecto específico de la preservación digital.
3. *Implantación del modelo óptimo de acceso y difusión del patrimonio digital.* Detallándose el modelo, la arquitectura, la tecnología utilizada, los estándares y normativa a aplicar, etc.

El resultado de esta acción se ha concretado en el desarrollo y puesta en marcha de Ondarenet, el archivo electrónico del patrimonio digital vasco (www.euskadi.net/ondarenet). Este archivo tiene la función de capturar, conservar y difundir el patrimonio que nace ya en formato digital y que trate sobre cualquier tema relacionado con la lengua, la sociedad o la cultura vasca.

Una vez definidos los objetivos, tareas y alcance del proyecto que fueron plasmados en una memoria de actuación, en 2007 se inició el proyecto por parte del Departamento de Cultura del Gobierno Vasco con el apoyo de la Sociedad Informática del Gobierno Vasco (EJIE).

3.1. Selección del soporte informático.

Se valoraron dos posibilidades: contratar un software comercial que desarrollara las herramientas necesarias a medida o utilizar alguno de los programas utilizados en proyectos similares.

Desde 1996 distintas bibliotecas nacionales han desarrollado programas propios para la recolección de sitios web. Es el caso de Suecia, país pionero a la hora de poner en marcha en 1996 el proyecto Kulturarw³ que pretende capturar la “web sueca” de manera integral. En un primer momento La Kung Royalbiblioteket utilizaba el software Combine, desarrollado por la Universidad de Lund y que permitía la recolección, tanto de elementos textuales, como imágenes y sonidos, aunque actualmente utiliza el robot de captura Heritrix. También la Biblioteca Nacional de Australia utiliza un software “ad hoc” denominado PANDAS (PANdora Digital Archiving System) implementado en 2001 y que permite, además de capturar urls, gestionar metadatos para su la descripción de los recursos capturados.

Tras valorar la posibilidad de desarrollar un software a medida par el proyecto Ondarenet, finalmente se optó por utilizar el Toolkit propuesto por el International Internet Preservation Consortium (IIPC) y que está siendo utilizado por un buen número de las iniciativas de archivo web internacionales existentes en este momento. Se trata de un conjunto de herramientas de código abierto que permiten la captura, gestión y visualización de los recursos capturados generando ficheros WARC, un formato que se ha convertido en estándar internacional mediante la reciente publicación de la norma ISO 28500:2009 Information and documentation -- WARC file format.

El toolkit está compuesto por un robot de captura que realiza el proceso de recolección de los componentes digitales -sitios y páginas web- de la colección denominado Heritrix, un motor de búsqueda de código abierto que permite la búsqueda e indización de los elementos recolectados basado en dos herramientas: NutchWAX y Wera y una aplicación que permite visualizar los resultados de la búsqueda Wayback. Finalmente, la programación y planificación de las capturas se llevan a cabo por medio de Web Curator Tool, una herramienta diseñada por la Biblioteca Nacional de Nueva Zelanda en colaboración con la British Library que permite gestionar de una manera más cómoda los procesos de captura y recolección de los elementos digitales (urls) que van a componer la colección.



Figura 1. Toolkit utilizado para la captura de webs en Ondarenet

3.2 El modelo selectivo

Decidir qué recursos capturar y conservar, ha sido uno de los puntos más importantes del proyecto. La mayor parte de las iniciativas relacionadas con el archivo de webs nacionales definen sus capturas según dos modelos. Por un lado existe el modelo integral o exhaustivo, cuyo principal exponente es Suecia, y que consiste en realizar una serie de “instantáneas” de toda la web de un país. Por otro lado se encuentra la captura selectiva, llevada a cabo por Australia, y que consiste en realizar capturas de las webs más representativas del país desde una política selectiva bien definida basada en criterios como la calidad, el tema, la lengua, etc.

En Ondarenet se ha optado por el modelo selectivo debido, principalmente, al hecho de que Euskadi no posee un dominio propio, lo que imposibilita realizar una captura exhaustiva de toda la web vasca. Por ello es imprescindible llevar a cabo una colección basada en una selección previa de los recursos a capturar. Esta opción permite crear colecciones equilibradas y facilita la difusión de los contenidos capturados (9).

Al hablar del universo digital de la Comunidad Autónoma de Euskadi se hace referencia tanto al conjunto de entidades e instituciones públicas o privadas e individuos productores de elementos digitales como páginas web -tanto estáticas como dinámicas-, recursos de comunicación como blogs, foros o listas de distribución y ficheros digitales asociados a los contenidos. En este sentido, en junio de 2009 se han identificado cerca de 800 recursos webs que están siendo capturados de manera progresiva, al mismo tiempo que se siguen identificando urls de interés utilizando métodos como la búsqueda directa a través de buscadores o aprovechando la propia hipertextualidad de la web que permite llegar a varios recursos de interés desde una url inicial.

3.3 Modelo de difusión

Desde un primer momento se planteó Ondarenet como un proyecto con dos objetivos bien diferenciados. Por un lado, se pretende preservar aquellos contenidos digitales accesibles por Internet relacionados con la vida política, social y cultural vasca y, por otro, garantizar el acceso a una información de interés que de otro modo se perdería. Es por ello que Ondarenet ofrece acceso a todos los recursos capturados y sus distintas versiones a través de una sencilla interfaz trilingüe (euskera, castellano e inglés) dirigida a todo tipo de usuario. Dicha interfaz permite llevar a cabo dos tipos de búsquedas: una búsqueda simple, que permite realizar búsquedas básicas bien por un término o términos concretos, bien por una url determinada, y una búsqueda avanzada a través de la cual se pueden restringir las búsquedas por fecha de captura, formato del recurso (imagen, pdf, etc.) o colección.

El usuario puede navegar además por un índice temático dividido en 12 categorías principales cada una de las cuales se subdivide en subcategorías más específicas lo que permite acceder a recursos capturados pertenecientes a una temática común.

Mediante la aplicación OndareNet podrá realizar búsquedas entre los sitios web capturados, archivados e indizados por el Servicio de Bibliotecas del Gobierno Vasco.

Formulario de búsquedas

Búsqueda simple

Texto:  

Url:  

Búsqueda avanzada

Texto:  

Formato:  Desde:  Hasta: 

Colección:  Orden: 

- Arte
- Cultura
- Educación e investigación
- Euskera
- Política y gobierno
- Sociedad
- Hechos relevantes
- Ciencia y tecnología
- Economía y negocios
- Empresa
- Ocio y cultura
- Salud
- Sociedad de la información

Figura 2. Interfaz de búsqueda de Ondarenet

4. Planteamientos a corto plazo

Es evidente que preservar la memoria histórica de un país pasa inevitablemente por preservar su memoria digital. Día a día se publican en la red contenidos e informaciones que serán, sin duda alguna, el punto de partida para los investigadores en un futuro no muy lejano. De ahí que sea necesaria una política de preservación y almacenamiento de estos contenidos digitales, sin olvidar aspectos tan importantes como su difusión y acceso.

En este sentido, la implantación de Ondarenet se puede valorar positivamente, ya que desde el inicio se consideró como un proyecto estratégico, y ha contado en todo momento con la implicación de EJIE (Sociedad Informática del Gobierno Vasco). La colaboración con EJIE ha sido vital para la definición y puesta en marcha de este proyecto de preservación web, máxime tratándose de una aplicación de software libre, no contemplada en los estándares informáticos del propio Gobierno Vasco.

Desde octubre de 2008, fecha en la que se comenzaron a realizar capturas con regularidad, se han descargado e indexado 112 webs, lo que supone un volumen total de más de 83 Gb correspondientes a 46.543 ficheros, tareas para las que se ha invertido un tiempo de descarga de 344 horas. El objetivo es continuar con las descargas con el fin de realizar al menos una captura anual de cada una de las webs identificadas. Una vez realizada esta primera “fotografía” de la web se pretenden llevar a cabo capturas semestrales con el fin de ofrecer una visión óptima de lo que se puede denominar la web vasca

Paralelamente se considera fundamental realizar colecciones especiales de recursos relacionadas con acontecimientos relevantes para la vida política, social, cultural de Euskadi. Así, durante los meses de febrero a abril de 2009 se

ha conformado la primera colección especial dedicada a las Elecciones al Parlamento Vasco 2009 en la que se han capturado páginas webs de partidos políticos y blogs de candidatos, entre otros recursos, que permiten acceder a una valiosa información que ya no está accesible en la red.

Con Ondarenet Euskadi se suma a una serie de iniciativas internacionales encaminadas a preservar recursos webs nacionales. Por ello se pretende formar parte en breve del Internacional Internet Preservatium Consortium (IIPC), en el que ya participan la mayor parte de las instituciones que lideran proyectos relacionados con la preservación de recursos digitales y que promueve la colaboración internacional con el fin de fomentar el desarrollo y el uso de herramientas y normas para la creación de archivos digitales.

Mientras se avanza en la captura de las webs seleccionadas, se trabaja también en la definición y creación de un repositorio institucional OAI que permita la integrar en una única herramienta tanto las webs capturadas como los objetos digitales procedentes de la Biblioteca Digital Vasca, constituida, hasta el momento, por los fondos digitalizados de la Fundación Sancho el Sabio y del Parlamento Vasco.

Además, el repositorio digital único permitirá la descripción de los recursos mediante el uso de estándares internacionales como son Dublin Core y METS lo que favorecerá la interoperabilidad de los registros y fomentará el acceso al patrimonio digital vasco desde una única interfaz.

Esta serie de iniciativas proyectadas a corto y medio plazo están encaminadas a impulsar la preservación y difusión del patrimonio digital vasco. Se trata en definitiva de crear una importante colección de recursos digitales, imprescindible para el estudio de la sociedad vasca en cualquiera de sus facetas -política, sociológica, o cultural-, ofreciendo a la ciudadanía una gran selección de recursos digitales archivados, pero además, clasificados y accesibles en línea, desde Internet.

(1) DÍAZ, A. La edición científica tradicional frena la difusión del saber. En: Campus, N. 548 (22 de abril de 2009). Disponible en web <<http://www.elmundo.es/suplementos/campus/2009/548/1240351203.html>> [Consulta: 27 junio 2009]

(2) ESPAÑA. Ministerio de Industria Turismo y Comercio. Dominios. Disponible en web <<https://www.nic.es/index.action>> [Consulta: 25 junio 2009]

(3) ESPAÑA. Ministerio de Cultura. Estadísticas del Libro, Lectura y Letras. Disponible en web <<http://www.mcu.es/libro/IN/estadisticas/index.html>> [Consulta: 25 junio 2009]

(4) International experts meet to tackle global digital memory loss. British Library Press Room, sept. 2008. Disponible en web <<http://www.bl.uk/news/2008/pressrelease20080925.html>> [Consulta: 25 junio 2009]

(5) ORERA ORERA, L. Preservación digital y bibliotecas: un nuevo escenario. En: Revista General de Información y Documentación, n.18 (2008) pp. 9-24. Disponible en web

<<http://www.ucm.es/BUCM/revistas/byd/11321873/articulos/RGID0808110009A.pdf>>

[Consulta: 25 junio 2009]

(6) UNESCO. Carta para la preservación del patrimonio digital. Disponible en web

<http://portal.unesco.org/ci/en/files/13367/10676067825Charter_es.pdf/Charter_es.pdf>

[Consulta: 26 junio 2009]

(7) EUSKADI. Ley 11/2007, de 26 de octubre, de Bibliotecas de Euskadi. Disponible en web <

[http://www.kultura.ejgv.euskadi.net/r46-](http://www.kultura.ejgv.euskadi.net/r46-4879/es/contenidos/informacion/recursos_profesional/es_recursos/adjuntos/leybib071026.pdf)

[4879/es/contenidos/informacion/recursos_profesional/es_recursos/adjuntos/leybib071026.pdf](http://www.kultura.ejgv.euskadi.net/r46-4879/es/contenidos/informacion/recursos_profesional/es_recursos/adjuntos/leybib071026.pdf)> [Consulta: 26 junio 2009]

(8) EUSKADI. Departamento de Cultura. Plan Director de digitalización, preservación y difusión del Patrimonio Cultural Vasco. Disponible en web

<[http://www.kultura.ejgv.euskadi.net/r46-](http://www.kultura.ejgv.euskadi.net/r46-19803/es/contenidos/informacion/keb_publicaciones/es_publicac/adjuntos/PatrimonioDigitalVasco_es.pdf)

[19803/es/contenidos/informacion/keb_publicaciones/es_publicac/adjuntos/PatrimonioDigitalVasco_es.pdf](http://www.kultura.ejgv.euskadi.net/r46-19803/es/contenidos/informacion/keb_publicaciones/es_publicac/adjuntos/PatrimonioDigitalVasco_es.pdf)> [consulta: 30 de junio 2009]

(9) LLUECA, C. Webs siempre accesibles: las bibliotecas nacionales y los depósitos

digitales nacionales. En BID: textos universitarios de biblioteconomía i documentació, n.

15 (2005). Disponible en web < <http://www.ub.es/bid/pdf/15lluec2.pdf>> [Consulta: 26

junio 2009]