Universidad del País Vasco, 29 y 30 abril 2009
II Jornadas sobre Documentación y Gestión de los contenidos digitales
http://www.doygestion.com/wp-content/uploads/2009/04/diptico.pdf

La experiencia catalana archivando la red: el repositorio Padicat (Patrimonio Digital de Cataluña) de la Biblioteca de Catalunya.

Ciro Llueca
Daniel Cócera

PADICAT (Patrimonio Digital de Cataluña)
Biblioteca de Catalunya
Hospital, 56
08001 Barcelona
www.padicat.cat

Resumen:

Objetivos y metodología

Como expone la Unesco en sus *Directrices para la preservación del patrimonio digital*, los recursos que son fruto del conocimiento o la expresión de los seres humanos, ya sean de carácter cultural, educativo, científico o administrativo, o comprendan información técnica, jurídica, médica o de otro tipo, se generan cada vez más a menudo directamente en formato digital, o se convierten a este formato a partir de material analógico ya existente.

Ello porque las tecnologías de la información y la comunicación han propiciado un crecimiento exponencial de la información de todo tipo y temática publicada en formato digital, en Internet. Pero por la propia naturaleza dinámica de la Red, gran parte de esa información muta a diario: es sustituida o simplemente desaparece. Es decir, la información producida en formato digital, contenida en las páginas web, corre el riesgo de colársenos por las manos como arena fina, si no se instrumentalizan proyectos de preservación y conservación. El reto y también el objetivo es construir repositorios que permitan acceder de manera permanente a los contenidos publicados en Internet.

A tal efecto y desde 1996, administraciones de diversos países han iniciado esos proyectos, llamados mayoritariamente “archivos Web”, con la meta de garantizar el

libre acceso a lo que se considera el patrimonio digital nacional. Suecia (proyecto *Kulturarw3*), Australia (proyecto *Pandora*), y el ambicioso proyecto *Internet Archive*, de alcance internacional, fueron los pioneros, y siguieron con el tiempo otros países como Gran Bretaña, Dinamarca, o Japón.

En España, la Biblioteca de Catalunya creó el repositorio Padicat (Patrimonio Digital de Cataluña), que ha afirmado su presencia internacional gracias a su inclusión en el International Internet Preservation Consortium (IIPC).

La Biblioteca de Catalunya, cuya misión es recopilar, conservar y difundir la producción bibliográfica catalana, considera patrimonio digital nacional toda aquella publicación en formato digital orientada, en el sentido más amplio, al público de Cataluña. La estrategia de preservación digital se concreta en las páginas web publicadas en lengua catalana u otras lenguas, bajo el dominio .CAT y otros dominios geográficos o temáticos, que estén relacionadas temáticamente con Cataluña.

Así nació en junio del año 2005 el repositorio Padicat, con un presupuesto aproximado de un millón de euros y el objetivo de crear la bibliografía digital catalana, el archivo web de Cataluña, siendo una iniciativa pionera en las comunidades de habla hispana, y con vocación de impulso en la futura e ineludible realización del patrimonio digital del resto de bibliotecas españolas.

Siguiendo la pauta generalizada en la mayoría de bibliotecas nacionales que se han sumado a este tipo de proyectos, se optó por la vía híbrida de captura de recursos web. Eso es, en primer lugar, una captura exhaustiva y automatizada de recursos digitales publicados en Internet realizada por el programa informático Heritrix. En segundo lugar, los acuerdos con una amplia selección de los sitios web representativos del entramado que conforma la sociedad civil catalana, tales como empresas, asociaciones profesionales, culturales o deportivas, partidos políticos y sindicatos, universidades, ayuntamientos, etc. En tercer lugar, la promoción de determinadas líneas de investigación futura mediante la integración focalizada de los recursos digitales de determinados acontecimientos, como han sido las campañas electorales en Internet para

las elecciones al Parlamento de Cataluña de 2006, municipales de 2007, al Congreso y Senado en 2008, o Europeas 2009.

Resultados

Tras tres años de Padicat, las anteriores acciones han producido un resultado que permite a la BC ofrecer en abierto, en línea, un total de unos 5.000 recursos digitales capturados en base a la captura sistemática y por los acuerdos con más de 400 instituciones, así como las referidas campañas electorales en Internet, y lo más importante, la necesaria preparación para asegurar con garantías el acceso permanente a esa parte de la Web catalana.

En el lado mejorable de la iniciativa, la optimización de la gestión relativa a sumar acuerdos de cooperación con las instituciones con el objetivo para el año 2011, que pretende abarcar el millar de organizaciones; y el desarrollo del programario dedicado a la captura y gestión de los recursos digitales publicados en Internet necesita aún de mucha dedicación técnica para optimizar sus resultados.

Conclusiones

Los beneficios de un repositorio como Padicat llegan a todos los sectores de la sociedad: para la ciudadanía representa el acceso abierto y permanente a los recursos que son fruto del conocimiento y expresión de los creadores del siglo XXI en un territorio determinado. Para las instituciones, empresas, administraciones y particulares que producen páginas web en ese territorio, la preservación de la propia producción y garantía de acceso, con los condicionantes que en cada caso la ley regula, a los contenidos y diseños que, de otro modo, desaparecerían. Y para el sistema bibliotecario, posibilidades infinitas de cooperación con el resto de bibliotecas, archivos y museos.

Palabras clave:

Archivos web; repositorios digitales; preservación digital; bibliotecas digitales

Breve currículo de los autores:

Ciro Llueca es coordinador del Padicat (Patrimonio Digital de Cataluña) en la Biblioteca de Catalunya. Anteriormente, fue responsable del área de Documentación del Centre de Cultura Contemporánea de Barcelona (CCCB). Diplomado en Biblioteconomía y Documentación por la Universitat de Barcelona, Licenciado en Documentación por la Universitat Oberta de Catalunya, y Master en Documentación Digital por la Universitat Pompeu Fabra. Profesor de la Facultad de Documentación de la UB, y consultor de los estudios de Documentación de la UOC. Fue vicepresidente (2003-2006) del Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya. Es concejal (2007-2011) de Cultura en el ayuntamiento de Figueres (Girona). Autor de "Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales" (BID 15, 2005), y "Archivando la Web, el proyecto Padicat" (El profesional de la información 15, 2006).

Daniel Cócera es responsable de gestión de colecciones del Padicat (Patrimonio Digital de Cataluña) en la Biblioteca de Catalunya. Licenciado en Historia por la Universitat Autònoma de Barcelona, y en Documentación por la misma universidad. Es coautor de "Padicat: realitat i reptes de 3 anys de l'arxiu web de Catalunya" (11es Jornades Catalanes d'Informació i Documentació, 2008).

TEXTO COMPLETO

La experiencia catalana archivando la red: el repositorio Padicat (Patrimonio Digital de Cataluña) de la Biblioteca de Catalunya.

1. Introducción: ¿Qué es el patrimonio digital?

Las tecnologías de la información y la comunicación han permitido que el patrimonio cultural y científico se presente en formato digital. Tal como se expresa la Unesco en sus *Directrices para la preservación del patrimonio digital*[i], los recursos que son fruto del conocimiento o la expresión de los seres humanos, ya sean de carácter cultural, educativo, científico o administrativo, o comprendan información técnica, jurídica, médica o de otro tipo, se generan cada vez más a menudo directamente en formato digital, o se convierten a este formato a partir de material analógico ya existente.

Pero ya antes de la publicación de estas Directrices, desde aproximadamente el inicio de la década de los 90 (1992-1995), el auge en progresión geométrica de los servidores y páginas web de Internet, motivaron la necesidad de articulación de proyectos para la creación de repositorios con el objetivo de preservar totalmente o en parte la producción documental sustentada en las páginas web, publicada en Internet.

Así, desde 1996 las administraciones de diversos países han llevado a cabo estrategias para garantizar el acceso permanente a la producción digital propia. Estas acciones están orientadas a asegurar en la medida de las posibilidades tecnológicas actuales a la adecuación del ciclo documental clásico a las páginas web: la compilación, el procesamiento, la preservación y el acceso permanente a la producción bibliográfica digital.

El reto no es menor y las amenazas son múltiples: por una parte, la manifiesta obsolescencia del texto legal que posibilita el depósito legal en España. Complementariamente, el crecimiento exponencial de la producción digital, sumado a la baja permanencia de los materiales publicados en Internet y, por supuesto y finalmente, el respeto a la legislación en materia de propiedad intelectual.

La propia naturaleza dinámica de la red es el factor de máxima erosión de los documentos que alberga. La información muta a diario: es sustituida o simplemente desaparece. El diseño, el entorno físico, evoluciona sin dejar rastro del anterior. Los nombres de los dominios varían, haciendo que, de un día para otro, no localicemos un recurso que ayer estábamos consultando. Como arena fina ente las manos desaparece a diario una ingente producción literaria, científica, educativa, lúdica o de creación artística, una producción que debemos considerar patrimonial, más aún cuando, como se expone en las referidas directrices, la vía digital es y será una de las principales herramientas, bien de creación, bien de difusión, de los creadores del siglo XXI.

Pese a estas dificultades, diversos países están realizando acciones de preservación de la producción digital más obvia: las páginas web. Las bibliotecas nacionales han sido a menudo impulsoras de estas acciones y en el caso español, la Biblioteca de Catalunya (BC) puso en marcha en 2005 el repositorio Padicat (Patrimonio Digital de Cataluña)[ii], dedicado al archivo sistemático de la producción digital en Cataluña.

2. El mundo archiva la Web

El archivo de la Web, como popularmente se conoce al conjunto de técnicas dirigidas para la creación de repositorios digitales como Padicat, no es hoy en día una acción consolidada en la mayoría de bibliotecas nacionales o entre otros órganos competentes en preservación patrimonial.

Un repositorio (o depósito) digital nacional es la herramienta fruto de la iniciativa dedicada a compilar, procesar y dar acceso a los recursos digitales de todo tipo creados en un territorio determinado, o sobre este territorio. Estos “archivos web” son

complementarios a los depósitos institucionales (como el *E-Prints* (*http://eprints.ucm.es/*) de la Universidad Complutense de Madrid o el *DSPace Revistes* y el *DSPace Eprints* (*https://upcommons.upc.edu/)* de la Universitat Politècnica de Catalunya, entre muchas otras iniciativas); o temáticos (como la *Biblioteca Virtual Miguel de Cervantes* (*http://www.cervantesvirtual.com)*, el portal *Tecnociencia e-revistas* (*http://www.tecnociencia.es/e-revistas)*, o el portal *Temaria* (*http://temaria.net*) que coordina la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona.

Existen diversos repositorios nacionales en funcionamiento, así como extensa bibliografía que los ha detallado y analizado[iii]. Los más conocidos son también los que dieron los primeros pasos en 1996: el sueco *Kulturarw3* (*http://www.kb.se/english/find/internet/websites/)* y el australiano *Pandora* (*http://pandora.nla.gov.au/index.html*); así como un repositorio de alcance internacional, el gigante *Internet Archive* (*http://www.archive.org*). Trece años más tarde podemos contar hasta 36 proyectos en diversas fases de implementación, siendo acciones consolidadas un tercio de esta cifra.

El análisis de estas experiencias muestra dos modelos básicos de sistemas con una tendencia generalizada hacia un modelo híbrido: El primero es el modelo integral o exhaustivo (mayoritario, y característico especialmente de los países escandinavos), que persigue la integración automática de la Web a partir de determinados criterios infraestructurales (lingüísticos, según el dominio de las páginas web, según la ubicación del servidor, etc.). El segundo modelo es el selectivo (asimilado por Australia, el Reino Unido o Japón, entre otros países), dirigido a compilar la Web en base a una política selectiva (sobre un espacio geográfico determinado, un tema de interés nacional, etc.). Estos dos modelos han dado paso, en lo que es ya una tendencia generalizada, a modelos híbridos (cuya caso más evidente es el de Dinamarca) que complementan la captura periódica de la Web con acciones selectivas, y en ciertos casos ampliando además esa cobertura a determinados acontecimientos de interés social (elecciones, competiciones deportivas, etc.).

Lamentablemente, el número de depósitos que, como Padicat u *Ondarenet* (*http://www.ondarenet.kultura.ejgv.euskadi.net/*), permiten acceder libremente a sus colecciones o a su fondo, es muy limitado. A menudo se trata de evitar potenciales conflictos con la vulneración de los derechos de propiedad intelectual de los recursos capturados sin autorización expresa, aunque complementariamente a esta legítima preocupación, un factor determinante radica en el hecho que no se hayan perfeccionado las interfaces de visualización de la información depositada.

3. La Biblioteca de Catalunya y el repositorio Padicat

A las puertas de su centenario en 2007, la BC apostó con firmeza por una serie de acciones para evolucionar hacia un modelo de biblioteca abierta, fiable y orientada al usuario. Una de las líneas estratégicas ha sido el impulso de proyectos digitales[iv], de carácter eminentemente cooperativo, para contribuir a la preservación del patrimonio catalán y aumentar la presencia de contenidos catalanes en Internet. Algunos de esos proyectos son *ARCA* (Archivo de Revistas Catalanas Antiguas), *RACO* (Revistas Catalanas de Acceso Abierto), *CLACA* (Clásicos Catalanes), el proyecto *Google Books* de digitalización de parte del fondo de la Biblioteca libre de copyright, o el propio Padicat.

De hecho, tal como se expresa en las leyes catalanas de bibliotecas de 1981 y 1993[v], la BC tiene por misión recopilar, conservar y difundir la producción bibliográfica catalana y la relacionada con el ámbito lingüístico catalán, y velar por la conservación y la difusión del patrimonio bibliográfico. Entendemos que este patrimonio bibliográfico incluye también la producción bibliográfica digital, la que se publica en Internet.

Así, tomando la misión ya descrita e incluyendo en ella la producción digital, establecemos el objetivo genérico de Padicat: diseñar y producir un sistema que permita a la BC compilar, procesar y dar acceso permanente a la producción digital catalana, confeccionando de esta manera la bibliografía digital de Cataluña.

Padicat es un depósito digital pionero en España y cuenta con la colaboración del Centre de Supercomputació de Catalunya (Cesca) y de la Generalitat de Catalunya, mediante su *Secretaria de Telecomunicacions i Societat de la Informació*. Tiene un presupuesto aproximado cercano al millón de euros. Está basado en el modelo híbrido (modelo danés), que se vertebra alrededor de tres ejes: la compilación masiva de recursos digitales; la selección de recursos digitales representativos de la Internet catalana; la creación de centros temáticos de interés entorno a acontecimientos de la vida pública catalana.

El día 11 de septiembre de 2006 se inauguró la web de Padicat (www.padicat.cat) en una versión trilingüe que hoy, corregida y aumentada, se mantiene. Desde el primer día, como filosofía del repositorio, se ha dado acceso abierto vía Internet a toda la colección disponible. Primero, con un motor de búsqueda a texto completo. En una segunda fase, con la creación de centros de interés: paquetes temáticos de interés público como puedan ser las campañas electorales en Internet, que se analizan más profundamente en líneas venideras. Finalmente, se han completado las opciones anteriores con la opción de búsqueda a través de URL y, sobretodo, de un directorio temático, dedicado al público que prefiere la navegación como fórmula de visita del fondo que forma la colección de Padicat. Más allá del éxito de estos trabajos, a día de hoy debemos afirmar que los procesos de posicionamiento de los resultados en respuesta a las búsquedas por texto completo, y muy especialmente la presentación de los recursos digitales capturados, están aún lejos de poderse considerar como óptimos.

Pese a todo, el depósito Padicat ha ayudado a posicionar a la Biblioteca de Catalunya en una situación de liderazgo en lo referente a preservación digital de páginas web. A nivel internacional, el repositorio forma parte desde febrero de 2007 de la principal red de trabajo en preservación digital, el International Internet Preservation Consortium (IIPC), y ha sido distinguido en enero de 2008 por la Library of Congress, responsable de comunicación de este consorcio, como ejemplo de archivo web por sus acciones en la campaña electoral de las elecciones municipales 2007.

Por otro lado, la BC ha asistido en estos tres años a un centenar de actos profesionales para explicar la iniciativa, proyectando una imagen de liderazgo en preservación del patrimonio digital, y ha tenido un impacto permanente en medios de comunicación especializados, y también generalistas, gracias a la emisión periódica de comunicados de prensa y otras fórmulas informativas.

Esta vocación de liderazgo es la que ha motivado la estrategia de la BC, basada en el aprendizaje constante tomando como espejo a los líderes mundiales en preservación digital, las entidades internacionalmente pioneras, como el referido *Internet Archive*, las bibliotecas nacionales escandinavas, los grupos de trabajo de estos organismos, etc. Hoy sabemos que tanto la distancia física como la lengua de contacto, en todo caso, no permiten el aprovechamiento de las sinergias (proyectos idénticos con objetivos similares alrededor del mundo) en la medida de lo deseable. Las listas de distribución y las reuniones esporádicas no suplen cualitativamente, aún, las posibilidades de aprendizaje mutuo. La inexistencia de proyectos similares en España, hasta el impetuoso arranque de Ondarenet, en el País Vasco, no ha posibilitado compartir experiencias en un entorno que trabaja cooperativamente en otras materias comunes. Por otro lado, somos conscientes que los proyectos internacionales de depósito digital nacional que están consolidados, no dedican a estas tareas los recursos necesarios para la mejora permanente de sus herramientas, mejora de la que se podrían beneficiar proyectos como Padicat.

Como se apuntaba en la definición esquemática del repositorio, el modelo que ha adoptado la BC es el sistema híbrido, en la línea de la referida tendencia generalizada en las bibliotecas nacionales, consistente en compilar masivamente los recursos digitales publicados en abierto en Internet; impulsar los acuerdos selectivos con los agentes implicados en la producción digital en Cataluña; y promover líneas de investigación específicas por medio de la integración focalizada de recursos digitales sobre determinados acontecimientos de la vida pública catalana.

Ello se traduce, actualmente y después de cuatro años de experiencia, en cerca de 450 acuerdos de colaboración con entidades de todo tipo que conforman el entramado

cívico, cultural, empresarial, deportivo, etc., de la sociedad catalana: la mayoría de universidades catalanas, colegios y asociaciones profesionales, clubs y federaciones deportivas, empresas, ayuntamientos, partidos políticos, museos, centros culturales, medios de comunicación y un largo etcétera, representan los socios de Padicat.

Por otro lado, desde la puesta en funcionamiento de su portal web en el año 2006, Padicat ha apostado por el eje de trabajo que ha resultado ser el más impactante: la creación de líneas de investigación específicas por medio de la integración focalizada de recursos digitales sobre determinados acontecimientos de la vida pública catalana. Se ha procedido a elaborar recopilaciones especiales de recursos web en eventos como las elecciones autonómicas de 2006, las municipales de 2007, las elecciones generales de 2008, o las recientes elecciones al Parlamento Europeo, de 2009. Asimismo, abordando otros ámbitos de la vida pública fuera del estrictamente político, también se ha elaborado una sección monográfica basada en la música folk y tradicional catalana, en colaboración con la Escuela Superior de Música de Catalunya (ESMUC), y se prepara otra monográfico sobre museos y colecciones museísticas en Cataluña.

Estas iniciativas han tenido una muy buena aceptación por parte de los medios de comunicación, y muy especialmente de parte de los profesionales universitarios especializados en cada materia, a los que se ha conseguido "integrar" para asesorar a la BC en la identificación y selección de los recursos digitales a recopilar.

De los tres ejes mencionados anteriormente que conforman el sistema híbrido, quizá sea el relativo a la captura masiva de recursos publicados en abierto (y especialmente la captura de aquellas webs bajo dominio .CAT), el que aporta una de las pocas sombras en la realización de los objetivos iniciales del repositorio. En 2006, el proyecto firmó un acuerdo de colaboración con la Fundació puntCAT (administradora de los dominios .cat) para acceder a los 25.000 registros bajo este dominio. Pero más allá de diversas capturas parciales de estos registros, no se ha procedido, a octubre de 2009, a un proceso de captura masiva de estas páginas web, siendo la causa principal la capacidad limitada de los recursos destinados a captura y almacenamiento de las páginas web. La

mejora en este punto supone uno de los retos de futuro del repositorio, como se apuntará más adelante.

Demos ahora una mirada más focalizada a determinados aspectos de funcionamiento del sistema, para una mejor descripción del depósito Padicat.

4. **Funcionamiento del sistema**

El sistema se basa en el ciclo documental clásico de bibliotecas y servicios de información (compilación, proceso, difusión), y, aplazando un análisis somero del sistema informático que permite el ciclo para más adelante, identificamos los pilares del proceso en la captura de los recursos, la organización de los mismos, y el acceso permanente a la colección.

En lo que respecta a la captura de los recursos se llevó a cabo un ejercicio de definición del tipo de recursos digitales susceptibles de captura, así como del alcance temático del proyecto. Es evidente que la tecnología que se aplica a los sistemas de repositorio digital cambia y cambiará en el futuro de forma exponencial, por ello las variables sobre la naturaleza del recurso digital y el *software* utilizado dotan de diferente grado de complejidad a lo que conocemos como páginas web.

Sin entrar a valorar el verdadero núcleo informativo de un recurso digital (¿todo él? ¿sólo la portada? ¿algunos capítulos? ¿un vídeo incrustado en el recurso?), sí citaremos la definición usada habitualmente por los miembros del Laboratorio de Internet del CINDOC-CSIC**,** que servirá para definir qué entendemos genéricamente por recurso digital, o "página web"[vi]: Página web, o conjunto de páginas web ligadas jerárquicamente a una página principal, identificable por una URL y que forma una unidad documental reconocible e independiente de otras, bien por su temática, bien por su autoría, bien por su representatividad institucional.

Por tanto, entendemos que una web susceptible de formar parte de la colección del repositorio deberá cumplir dos condiciones básicas: será una página web identificable por una URL, y formará una unidad documental reconocible. No importa, pues, que ese recurso digital sea en un lenguaje de programación concreto, o el formato del recurso sea texto, imagen, sonido, etc. De hecho, ya las primeras pruebas del repositorio ofrecieron datos cristalinos sobre el elevado porcentaje de formatos estándares en las páginas web capturadas.

Por lo que respecta a la cobertura temática del repositorio Padicat, entendemos "Patrimonio Digital" como la información electrónica publicada en Internet, en abierto o no, independientemente del formato en que se presenta esta información. Entendemos "de Cataluña" en el sentido que tradicionalmente ha tenido la bibliografía nacional de Cataluña en que se basa la política de la BC: todo aquello producido en Cataluña, o que trate sobre Cataluña.

Lo cierto es que Internet está diseñada para romper barreras políticas y hacer la información accesible universalmente. Pese a este hecho definitorio, es posible identificar partes de esa red que contengan módulos de interés de grupos concretos, a los que podemos llamar "comunidades de usuarios web", referidas a cierta temática o simplemente de interés de una comunidad concreta. A efectos prácticos, se establece la estrategia de captura en: webs bajo dominio .CAT; o webs ubicadas en servidores de Cataluña; o webs bajo dominios geográficos (.ES, .COM, .NET, .ORG, etc.) en lengua catalana; o finalmente, webs que no cumplen los requisitos anteriores, pero relacionadas temáticamente con Cataluña.

5. El *software* de PADICAT

La organización de los recursos web, una vez capturados, debe permitir gestionar la colección y asegurar la recuperación, toda vez que ha de preservar los contenidos digitales con las técnicas disponibles. Esta organización incluye la identificación

permanente de los recursos, la aplicación de metadatos, el almacenamiento, y la preservación.

Tras la fase de análisis y testeo de programas, se determinó que el programa informático Heritrix[vii], usado por la mayoría de los proyectos como el que nos ocupa para la captura de recursos digitales, sería el utilizado por el sistema. Este programa es el encargado de recolectar las páginas web tal como las ve el usuario que navega por Internet, y almacenarlas en archivos comprimidos en formato ARC[viii]. Después, NutchWax[ix] y Hadoop[x] realizan un proceso de indexación de la información recolectada que permitirá, posteriormente, utilizar estos índices para localizar recursos dentro de la colección.

Existen dos interfaces para realizar las consultas al conjunto de recursos capturados: Wera[xi], que permite la búsqueda por palabras clave a través de los índices generados por NutchWax; y Wayback[xii], que permite la consulta directa por URL.

El programa Web Curator Tool[xiii], desarrollado por la National Library of New Zealand, se ha aprovechado como sistema de gestión documental que permite la asignación de metadatos a una parte significativa de la colección, lo que garantiza la posibilidad futura de integrar la colección en otros catálogos de la BC u otras instituciones (aunque para ello, como veremos, será necesaria una evolución considerable de las actuales prestaciones informáticas).

Todo el *software* que utiliza Padicat es de código abierto y gratuito y ha sido desarrollado por organizaciones sin ánimo de lucro asociadas al International Internet Preservation Consortium (IIPC).

Por otra parte, el sistema prevé un depósito que permita conservar todos los recursos, de manera que se tenga acceso en todo momento. Se contempla un sistema de doble copia en diferente ubicación geográfica, con una capacidad total necesaria de 20 TB en los períodos de producción y explotación del repositorio, hasta 2011.

El método de preservación merece un tratamiento en comunicaciones adicionales al presente artículo, pero somos conscientes de la problemática de las estrategias más

habituales de preservación[xiv], como la migración periódica o *refresh* de los datos (migración a nuevas versiones de los mismos programas o lenguajes, o a nuevos programas capaces de leer los anteriores), la emulación (el uso de *software*, especificaciones, etc., utilizado en el momento de la creación), la recreación (simulación por ingeniería inversa u otros métodos).

Y es que las plataformas tecnológicas utilizadas para este tipo de depósitos están muy centradas en los tres aspectos básicos de la cadena documental: la captura de recursos; la indexación de estos recursos una vez capturados; y el acceso a los recursos almacenados. Pero en menor medida en la preservación de estos recursos, territorio prácticamente virgen si lo asociamos exclusivamente a los procesos dirigidos a garantizar el acceso permanente a las páginas web capturadas y no, como sucede en la mayoría de casos, si se analiza des de una óptica más global (preservación de materiales digitalizados, preservación de ficheros ofimáticos, etc.).

En todo caso, las previsiones sobre el tipo de archivos que el repositorio debe gestionar, basadas en la actual composición del fondo de la colección, revelan que la mayor parte de los archivos corresponden a formatos estándares, que pueden simplificar la tarea preservadora al menos en las macrocifras. Así, sobre una muestra cercana a los 107 millones de ficheros, aproximadamente el 69% de ellos corresponden a formatos estándares, lo que significa el 87% del espacio ocupado actualmente: texto/html (65%), imagen jpeg o gif (4%), etc.

En cuanto al *hardware* utilizado, creemos que una somera descripción de las máquinas será suficiente para dar por vista la vertiente más física del repositorio: Padicat tiene a su disposición siete nodos HP ProLiant DL360 G4p encargados de las tareas de recolección e indexación de webs. De la búsqueda y visualización de resultados en la interfaz web se encarga un *clúster* Linux de alta disponibilidad con características de balanceo de carga de peticiones y de tolerancia a fallos en caso de desastre en los nodos que componen la plataforma. Los nodos están conectados mediante fibra a una *Storage Area Network* (SAN) y el sistema se completa con un robot donde se guardan copias de seguridad de los datos en cinta.

6. **Retos para el futuro**

El futuro de Padicat, después de una etapa que podemos considerar de nacimiento, pasa por consolidar su capacidad de crecimiento, mejorar sus procesos de trabajo y optimizar los recursos de que dispone.

En primer lugar, hace falta dimensionar la infraestructura necesaria del repositorio, adaptándola a los objetivos del sistema, o bien modificar a la baja estos objetivos. La actual estructura de *hardware* y de personal experto en el *software* utilizado no permite trabajar con la capacidad necesaria para acometer el reto de la captura global de la web catalana. El hecho de tener al CESCA como socio tecnológico sin duda deberá permitir establecer cuáles son las necesidades, y en base a estas poder dar una respuesta tecnológica para el crecimiento exponencial que perseguimos.

En segundo lugar, es imprescindible abordar la definición de las estrategias de preservación de los ficheros que contiene el repositorio. Probablemente sea este uno de los aspectos clave en el retorno que la BC quiere ofrecer a la sociedad. Al margen de radiografías periódicas de la web catalana, que ilustran el diagnóstico del lenguaje de programación usado en la edición digital, el sistema puede ayudar a definir cuáles son los formatos que a corto plazo sufren problemas de ilegibilidad. Una vez constatadas estas pérdidas, es posible identificar hacia qué formatos hace falta transformar los ficheros para dotarlos de una permanencia más longeva, así como los procesos que han de permitir esta transformación.

En tercer lugar, Padicat ha de seguir apostando por potenciar los ejes de trabajo que hasta la fecha han dado los mejores resultados, como son las recopilaciones especiales sobre eventos de la vida pública catalana expresados a través de Internet, y aprovechar estos centros de interés para la implicación de los colectivos expertos en el asesoramiento a la BC de los yacimientos de recursos digitales.

En cuarto lugar, el abordaje de la captura sistematizada de publicaciones en serie en Internet es un reto de futuro, que se iniciará en los próximos meses con capturas que permitirán proyectar las necesidades infraestructurales del repositorio. La revisión del *software* existente, para posibilitar el aprovechamiento de los ficheros ya descargados en sucesivas capturas cuando estas se repitan con mucha frecuencia, será la solución a este reto, porque será también la manera de optimizar espacio en disco, tiempo de captura, y, en definitiva, los recursos existentes.

Finalmente, pese a la estandarización de los lenguajes informáticos que se utilizan en el *software* de Padicat y el resto de proyectos similares, hace falta destacar que no es aún posible, como cabía esperar, un intercambio eficaz de registros bibliográficos, con la finalidad de poder integrar todos los depósitos existentes, o estos depósitos en otros catálogos. El uso de pasarelas y lenguajes estándares está aún en fase de implementación en el *software* del repositorio que, insistimos, es común al de la mayoría de depósitos digitales existentes en la red. De la capacidad de incidir en el desarrollo del *software* depende también la consecución de los objetivos de futuro de la BC, en su voluntad de archivar la Web catalana.

En la presentación de los objetivos iniciales de Padicat, en 2005, se planteaban los potenciales beneficios de un proyecto que se encontraba en un estadio preliminar. Cuatro años más tarde, los beneficios son plenamente vigentes desde el momento en que han llegado a ser factores críticos de éxito en la estrategia de la BC: a ojos de la comunidad bibliotecaria de Cataluña, los beneficios se centran en la integración de los documentos nacidos digitales en la bibliografía nacional, y en el contundente posicionamiento de la Biblioteca de Catalunya y sus socios de repositorio en una situación privilegiada como fuente de información de los documentos que representan, en buena medida, el futuro. Posibilidades infinitas de cooperación con las instituciones de la memoria, bibliotecas, archivos y museos de Cataluña, así como universidades y centros de investigación, e impulso y liderazgo en la confección del patrimonio digital de España. Así mismo, la ya existente relación de privilegio con el resto de bibliotecas nacionales del mundo en términos de preservación digital y depósitos nacionales. Para las instituciones, empresas, administraciones y particulares que producen páginas web

en Cataluña, preservación de la propia producción y garantía de acceso, con los condicionantes que rige la ley, a los contenidos y diseños que, de otra forma, desaparecerían. Y por último, para la ciudadanía, y como se pretende en las Directrices de la Unesco, acceso abierto y permanente a los recursos que forman el Patrimonio Digital de Cataluña.

7. **Bibliografía**

Biblioteca de Catalunya. *Memòria del plantejament del projecte PADICAT (Patrimoni Digital de Catalunya)*. Barcelona: Biblioteca de Catalunya, 2005. <http://www.recercat.net/handle/2072/1757> [Consulta: 30/09/2009]

Biblioteques digitals i dipòsits nacionals de recursos digitals. Barcelona: Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, 1999.

Cócera, D.; Llueca, C. (2008). "Padicat: realitat i reptes de 3 anys de l'arxiu web de Catalunya". *11es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya. <http://eprints.rclis.org/archive/00013562/01/llueca_padicat_jornades_2008.pdf>. [Consulta: 30/09/2009]

Directrices para la preservación del patrimonio digital. Canberra: Unesco, 2003. <http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>. [Consulta: 30/09/2009]

Gomes, D.; Silva, M. J. (2005). "Characterizing a National Community Web". *ACM Transactions on Internet Technology*, vol 5, num 3 (Aug 2005). <http://xldb.fc.ul.pt/daniel/gomesCharacterizing.pdf>. [Consulta: 30/09/2009]

Hodge, G. M. (2000). "Best practices for digital archiving: an information life cycle approach". *D-LIB Magazine*. Vol. 6, num. 1 (jan 2000). <http://www.dlib.org/dlib/january00/01hodge.html>. [Consulta: 30/09/2009]

Keefer, A.; Gallart, N. (2007). *La preservació de recursos digitals: el repte per a les biblioteques del segle XXI*. Barcelona: UOC.

Llueca, C. (2005). "Webs siempre accesibles : las bibliotecas nacionales y los depóditos digitales nacionales". *BiD: textos universitaris de biblioteconomia i documentació*, núm. 15 (des 2005). <http://www.ub.es/bid/15lluec2.htm> [Consulta: 30/09/2009]

Llueca, C. (2006). "El projecte PADICAT (Patrimoni Digital de Catalunya) de la Biblioteca de Catalunya", *10es Jornades Catalanes d'Informació i Documentació*. Barcelona: Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya. <http://eprints.rclis.org/archive/00006434/01/llueca_padicat.pdf>. [Consulta: 30/09/2009]

Llueca, C. (2006). "Archivando la Web, el proyecto Padicat (Patrimonio Digital de Cataluña)". *El profesional de la información*. Vol. 15, núm. 6, p. 473-478. <http://eprints.rclis.org/archive/00007767/01/epi_padicat.pdf> [Consulta: 30/09/2009]

Llueca, C. (2007). "Archivando la web catalana, el proyecto PADICAT", *Clip: Boletín de la SEDIC*. Núm. 47. <http://www.sedic.es/p_boletinclip47_confirma.htm>. [Consulta: 30/09/2009]

Torres, N; y otros (2007). "Patrimoni Digital de Catalunya, experiencias del primer año", *Jornadas Técnicas RedIris*. Oviedo: RedIris. <http://www.padicat.cat/docs/poster_padicat_rediris.pdf> [Consulta: 30/09/2009]

[i] *Directrices para la preservación del patrimonio digital*. Canberra: Unesco, 2003. [Consulta: 30/09/2009] *http://unesdoc.unesco.org/images/0013/001300/130071s.pdf*

[ii] El portal PADICAT, http://www.padicat.cat está operativo desde septiembre de 2006.

[iii] Para un panorámica gobal sobre estos proyectos véase: Llueca, Ciro. "Webs siempre accesibles: las bibliotecas nacionales y los depósitos digitales nacionales". En: *BiD: textos universitaris de biblioteconomia i documentació*, 2005, diciembre, n. 15. [Consulta: 30/09/2009] *http://www.ub.es/bid/15lluec2.htm*

[iv] La estrategia y los proyectos, a excepción del Google Books que fue un acuerdo posterior, fueron presentados en: Lamarca, D.; Serra, E. "L'estratègia de la Biblioteca de Catalunya en projectes digitals". En: *Ítem*, 2005, setembre-desembre, n. 41, pp. 41-43.

[v] Artículo 7.1 de la "Llei de biblioteques de Catalunya, de 24 d'abril de 1981". En: *Diari Oficial de la Generalitat de Catalunya*, 1981, 29 abril, n. 123. Dicho artículo se refrendó en la "Llei 4/1993 del sistema bibliotecari de Catalunya, de 18 de març de 1993". En: *Diari Oficial de la Generalitat de Catalunya*, 1993, 29 març, n. 1727.

[vi] Interesante reflexión terminológica en: Pareja, Víctor Manuel [et al.]. "Desarrollo y aplicación del concepto de sede web como unidad documental de análisis en Cibermetría". En: *9as Jornadas Españolas de Documentación*, 2005. La definición "oficial" del World Wide Web Consortium (*http://www.w3.org/2003/glossary)* define en su glosario el concepto "página web" como una colección de información consistente en uno o más recursos Web, pensados para ser utilizados simultáneamente, e identificados con un único URI.

[vii] Heritrix (*http://crawler.archive.org/)*. Un artículo fundamental, a cargo de uno de sus creadores, es Mohr, G. [et al.]. "An introduction to Heritrix: an open source archival quality web crawler". En: *International Web Archiving Workshop*, 2004. [Consulta: 30/09/2009] *http://www.iwaw.net/04/Mohr.pdf*

[viii] Arc File Format (http://en.wikipedia.org/wiki/ARC_(file_format))

[ix] NutchWax (http://archive-access.sourceforge.net/projects/nutch/)

[x] Hadoop (http://hadoop.apache.org/core/)

[xi] Wera (http://archive-access.sourceforge.net/projects/wera/)

[xii] Wayback (http://www.archive.org/web/web.php)

[xiii] Web Curator Tool (http://webcurator.sourceforge.net/)

[xiv] Ayre, Catherine; Muir, Adrienne. "The right to preserve: the rights issues of digital preservation". En: *D-Lib magazine*, 2004, march, v. 10, n. 3. [Consulta: 30/09/2009] *http://www.dlib.org/dlib/march04/ayre/03ayre.html*